(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 December 2000 (21.12.2000)

PCT

(10) International Publication Number
WO 00/077712 A1

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/US00/16638

(22) International Filing Date: 15 June 2000 (15.06.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/333,726 15 June 1999 (15.06.1999) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 09/333,726 (CIP)
Filed on 15 June 1999 (15.06.1999)

(71) Applicant *(for all designated States except US)*: **ROCKEFELLER UNIVERSITY** [US/US]; 1230 York Avenue, New York, NY 11021-6399 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ERIKSSON, Jan** [SE/SE]; Studentvagen 15, S-752 34 Uppsala (SE). **FENYÖ, David** [SE/US]; 104 Sullivan #18, New York, NY 10012 (US). **CHAIT, Brian, T.** [ZA/US]; 500 East 63rd Street #20D, New York, NY 10021 (US).

(74) Agent: **FEIT, Irving, N.**; Hoffmann & Baron LLP, 6900 Jericho Turnpike, Syosset, NY 11791 (US).

(81) Designated States *(national)*: CA, JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *with international search report*

(48) Date of publication of this corrected version:
11 July 2002

(15) Information about Correction:
see PCT Gazette No. 28/2002 of 11 July 2002, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 00/077712 A1

(54) Title: METHOD FOR ASSESSING SIGNIFICANCE OF PROTEIN IDENTIFICATION

(57) Abstract: A method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules, the method comprising: a) generating mass data for the particular experimental condition for known biological molecules in a biological molecule database; b) generating mass data of a hypothetical biological molecule for the experimental condition; c) comparing the data generated in step (b) with the data generated for each known biological molecule in step (a); d) calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data generated in step (a), which corresponds to a particular known biological molecule, and the data generated in step (b); e) selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the data generated in step (a) and the data generated in step (b); f) repeating steps (b) through (e) with different hypothetical biological molecules until a sufficient quantity of scores are selected; and g) determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

METHOD FOR ASSESSING SIGNIFICANCE OF PROTEIN IDENTIFICATION

BACKGROUND

An unknown biological molecule can be identified by comparing the mass data of the unknown biological molecule with mass data of known biological molecules.

For example, the rapid growth of available high quality DNA sequence data has made mass spectrometry (MS) combined with genome database searching a popular and potentially accurate method to identify proteins. Protein identification by mass spectrometry has proven to be a powerful tool to elucidate biological function and to find the composition of protein complexes and entire organelles.

In protein identification experiments, proteins are typically separated by gel electrophoresis, subjected to a protease having high digestion specificity (e.g. trypsin) and the resulting mixture of peptides is extracted from the gel and subjected to MS-analysis (1998). The distribution of proteolytic peptide masses (peptide map) is compared with theoretical proteolytical peptide masses calculated for each protein stored in a protein/DNA sequence database.

There are various algorithms that attempt to identify the protein with the highest degree of similarity to the experimentally obtained peptide map. These algorithms yield the protein identified and an identification score. Due to imperfections in the protein separation and to incomplete extraction of the proteolytic peptides from the gel, the peptide map is typically incomplete with respect to the protein identified, and also contains a background of proteolytic peptide masses from one or several other proteins. Even if separation and extraction were perfect, posttranslational modifications of proteins would cause a proteolytic peptide mass distribution different from that predicted by the genome. Mass spectrometry determines a peptide mass m_i to an accuracy $\pm\Delta m_i$, with $\Delta m_i/m_i$ typically >30ppm. Within the mass range $m_i \pm \Delta m_i$ proteolytic peptide masses of several proteins in the genome can match. For these reasons, a database search using the information in a

peptide map will not always identify a protein unambiguously.

Despite the momentum mass spectrometric protein identification has given to protein research, the problem of objectively assessing the significance of a protein identification result has been overlooked. As increasingly complex biological problems are explored, knowledge of the significance of each protein identification result is likely to become critical.

The object of the present invention is to provide a method for assessing the significance of a biological molecule identification.

SUMMARY OF THE INVENTION

This and other objects, as will be apparent to those having ordinary skill in the art, have been met by providing a method of determining the statistical significance of a biological molecule identification score. The method comprises a) selecting a significance level that represents a level of confidence in a biological molecule identification b) calculating a score associated with an unknown biological molecule, wherein the score is a function of similarity between mass data of the unknown biological molecule and mass data generated for known biological molecules of a biological molecule database; c) comparing the score with a score frequency distribution, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated for known biological molecules of a biological molecule database, and wherein the frequency distribution has associated therewith the significance level; and d) determining whether the score associated with the unknown biological molecule identification is within the significance level.

The invention further provides a method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules. The method comprises a) generating mass data for the particular experimental condition for known biological molecules in a biological molecule database; b) generating mass data of a hypothetical biological molecule for the experimental condition; c) comparing the data generated in step (b) with the data generated for each known biological molecule in step (a); d) calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data generated in step (a) which corresponds to a particular known biological molecule and the data generated in step (b); e) selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the data generated in step (a) and the data generated in step (b); f) repeating steps (b) through (e) with different hypothetical biological molecules until a sufficient quantity of scores are selected; and g) determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

The invention provides another method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules. The method comprises a) generating mass data for the particular experimental condition for known biological molecules in a biological molecule database; b) randomly selecting a biological molecule from the database; c) comparing the mass data of the randomly selected biological molecule with the mass data of each known biological molecule; d) calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data; e) selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a degree of similarity between the data which is lower than the highest degree of similarity; f) repeating steps (b) through (d) with different randomly selected biological molecules until a sufficient quantity of

scores are selected; and g) determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

The invention also provides a method of identifying an unknown biological molecule for a particular experimental condition and a particular significance level. The method comprises a) selecting a significance level that represents a level of confidence in a biological molecule identification; b) cleaving the unknown biological molecule into constituent parts by a method that produces constituent parts; c) generating mass data for these constituent parts; d) comparing the mass data generated in step (c) with mass data generated for the experimental condition from known biological molecules of a biological molecule database; e) calculating scores for each comparison in step (d), wherein the scores are a function of similarity between mass data of the unknown biological molecule and mass data generated from the biological molecule database; f) selecting a score generated in step (e) wherein the score corresponds to a comparison which denotes a high degree of similarity and wherein the score corresponds to a particular known biological molecule in the biological molecule database; and g) determining whether the score selected in step (f) is equal to or larger than the critical score.

In another embodiment the invention comprises a computer program product which comprises a computer usable medium having computer readable program code means embodied in said medium for generating a frequency distribution of scores, wherein the scores relate to random identifications of biological molecules. The computer program product includes: a computer readable program code means for causing a computer to generate mass data for each known biological molecule in a biological molecule database for a particular experimental condition; computer readable program code means for causing the computer to generate mass data of a hypothetical biological molecule for the experimental condition; computer readable program code means for causing the computer to compare the mass data of the

hypothetical biological molecule with the mass data generated for each known biological molecule in the biological molecule database for the particular experimental condition; computer readable program code means for causing the computer to calculate a score for each mass data comparison, wherein the score is a function of similarity between the mass data corresponding to a particular known biological molecule and the mass data corresponding to the hypothetical biological molecule; computer readable program code means for causing the computer to select a score from the calculated scores, wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the mass data corresponding to the particular known biological molecule and the mass data corresponding to the hypothetical biological molecule; computer readable program code means for causing the computer to repeatedly generate mass data of different hypothetical biological molecules, compare the mass data each of the hypothetical molecules with the mass data generated for each known biological molecule in the biological molecule database, calculate a score for each of the mass data comparisons and select a score from the calculated scores until a sufficient quantity of scores are selected; and computer readable program code means for causing the computer to determine the frequency of selecting each score and to generate therefrom a frequency distribution of scores.

In another embodiment the invention comprises a computer program product which comprises a computer usable medium having computer readable program code means embodied in said medium for identifying an unknown biological molecule for a particular experimental condition and a particular significance level. The computer program product includes: computer readable program code means for causing a computer to generate mass data of an unknown biological molecule, the unknown biological molecule having been cleaved into constituent parts by a method that produces constituent parts; computer readable program code means for causing the computer to compare the mass data of the unknown biological molecule with mass

data generated for the experimental condition from known biological molecules of a biological molecule database; computer readable program code means for causing the computer to calculate scores for each mass data comparison, wherein the scores are a function of similarity between mass data of the unknown biological molecule and mass data generated from the biological molecule database; computer readable program code means for causing the computer to select a score from the calculated scores, wherein the selected score corresponds to a particular known biological molecule in the biological molecule database, and wherein the selected score corresponds to a comparison which denotes a high degree of similarity; computer readable program code means for causing the computer to compare the selected score with a frequency distribution of scores for the experimental condition, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated from a biological molecule database, and wherein the frequency distribution has associated therewith a critical score which corresponds to the significance level; and computer readable program code means for causing the computer to determine whether the selected score is equal to or larger than the critical score.

In another embodiment the invention comprises a computer program product which comprises a computer usable medium having computer readable program code means embodied in said medium for determining statistical significance of a biological molecule identification score. The computer program product includes: a computer readable program code means for causing a computer to calculate a score associated with an unknown biological molecule, wherein the score is a function of similarity between mass data of the unknown biological molecule and mass data generated from a biological molecule database; computer readable program code means for causing the computer to compare the score with a score frequency distribution, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated from a biological molecule

database, and wherein the frequency distribution has associated therewith a significance level determined to represent a confident biological molecule identification; and computer readable program code means for causing the computer to determine whether the score associated with the unknown biological molecule identification is within the significance level.

DESCRIPTION OF FIGURES

Figure 1: Score frequencies for proteins identified from 300 different random peptides maps composed of 20 and 80 tryptic peptide masses respectively. a) Algorithm 1. b) Algorithm 2.

Figure 2: a) Frequency function f(S) for algorithm 1 assumed to describe the null hypothesis H_0 "the protein identification is random and false." The area α of the shaded region under f(S> and = S_C) represents the probability that a result for which H_0 is true has at least the score S_C. In significance testing α is called the significance level and is shown in b) as a function of S_C. c) Magnified portion of (b) with horizontal lines indicating the significance levels: 0.05 (*), 0.01 (**) and 0.001 (***). d-f) Parallels a-c, but with protein identification based on algorithm 2. The better resolution of the score variable in algorithm 1 makes it easier to determine accurately the score S_C that corresponds to a chosen α.

Figure 3: The scores corresponding to three different significance levels α =0.05 (*), 0.01 (**) and 0.001 (***) as a function of the number of tryptic peptide masses in a peptide map. The lines through the simulated data points represent least square fits to second order polynomial functions. a) Algorithm 1 b) Algorithm 2. In the inset, scores (number of matches) corresponding to α =0.01 are shown relative to the number of peptides in the maps.

Figure 4: The score S_C corresponding to α=0.01 as a function of the mass accuracy for random tryptic peptide maps with 20 and 50 masses respectively. Scores S_C were normalized to the respective values of S_C for a mass accuracy of 0.1 Da. Top panel: Algorithm 1 Bottom panel: Algorithm 2.

Figure 5: The protein identification score S_C required for 1% significance (α =0.01) versus u, the maximum number of uncleaved cleavage sites allowed in the database search for algorithm 1 (1) and algorithm 2 (2). The results were normalized to the respective values of S_C for u=0. a) random tryptic peptide maps with 20 masses b) Random tryptic peptide maps with 50 masses.

Figure 6: The score S_C that yields α =0.01 as a function of the maximum protein mass allowed in the database search for algorithm 1 (1) and algorithm 2 (2). The results are shown relative to the respective S_C obtained with a maximum protein mass of 100 kDa. a) Random tryptic peptide maps with 20 masses. b) Random tryptic peptide maps with 50 masses.

Figure 7: The influence of the size of the genome database on the protein identification score S_C required for 1% significance. The simulated data represent H. influenzae (h.i.), S. cerevisiae (yeast) and C. elegans (c.e.). The yeast genome was divided into two parts of similar size and the results from the respective parts were averaged ($y_{1/2}$). a) Algorithm 1 b) Algorithm 2.

Figure 8: Protein Identification Database Search with Information from a Peptide Map

Figure 9: Obtaining a Score Frequency Distribution.

Figure 10: Generation of a Hypothetical Peptide Map.

Figure 11: Protein Identification by MS Fragmentation.

Figure 12: Protein Identification by Peptide Mapping.

Figure 13: Flow Chart Showing the Steps in a complete protein identification MS/MS experiment.

Figure 14: Score distributions due to random matching from two different simulation models and algorithm 2. Similar results are obtained for the highest ranked protein when using random tryptic peptide maps (each mass from a different protein) as for the second highest ranked protein when using ideal tryptic peptide maps (all masses from a single protein). The latter method fails for large peptide maps.

DETAILED DESCRIPTION

In one embodiment the invention provides a method of identifying an unknown biological molecule. Biological molecules include any biological polymer that can be degraded into constituent parts. The degradation is preferably into constituent parts at predictable positions to form predictable masses. Examples of biological molecules include proteins, nucleic acid molecules, polysaccharides and carbohydrates.

Proteins are polymers of amino acids. Constituent parts of proteins comprise one or more amino acids. A protein typically contains approximately at least ten amino acids, preferably at least fifty amino acids and more preferably at least one hundred amino acids. A constituent part of a protein that contains more than one amino acid is referred to herein as a peptide.

Nucleic acids are polymers of nucleotides. Constituent parts of nucleic acids comprise one or more nucleotides. Typically, a nucleic acid contains approximately at least one hundred nucleotides, preferably at least five hundred nucleotides. A constituent part of a nucleic acid that contains more than one nucleotide is referred to herein as an oligonucleotide.

Polysaccharides are polymers of monosaccharides. Constituent parts of polysaccharides comprise one or more monosaccharides. Typically, a polysaccharide contains approximately at least five monosaccharides, preferably at least ten monosaccharides. A constituent part of a polysaccharide that contains more than one monosaccharide is referred to herein as an oligosaccharide.

Mass data of biological molecules is quantifiable information about the masses of the constituent parts of the biological molecule. Mass data includes individual mass spectra and groups of mass spectra. The mass spectra can be in the form of peptide maps, oglionucleotide maps or oligosaccharide maps.

Mass data for proteins can be generated in any manner which provides mass data within a certain accuracy. Examples include matrix-assisted laser desorption/ionization mass spectrometry, electrospray ionization mass spectrometry, chromatography and electrophoresis. Mass data can also be generated by a general purpose computer configured by software or otherwise.

For the purposes of the present invention the mass data, for example a peptide mass, m_i, is determined to an accuracy $\pm\Delta m_i$, with $\Delta m_i/m_i$ preferably <10,000 ppm, more preferably <100ppm and most preferably <30ppm.

A step in generating mass data of a biological molecule may include first

cleaving the biological molecule into constituent parts. Biological molecules may be cleaved by methods known in the art. Preferably, the biological molecules are cleaved into constituent parts at predictable positions to form predictable masses. Methods of cleaving include chemical degradation of the biological molecules. Biological molecules may be degraded by contacting the biological molecule with any chemical substance.

For example, proteins may be predictably degraded into peptides by means of cyanogen bromide and enzymes, such as trypsin, endoproteinase Asp-N, V8 protease, endoproteinase Arg-C, etc. Nucleic acids may be predictably degraded into constituent parts by means of restriction endonucleases, such as Eco RI, Sma I, BamH I, Hinc II, etc. Polysaccharides may be degraded into constituent parts by means of enzymes, such as maltase, amylase, alpha-mannosidase, etc.

The invention relates to improving current methods for identifying biological molecules by adding to current methods an assessment of the significance of the identification. Current methods for identifying biological molecules as well as the methods of the present invention will be described for protein identification. These methods are equally applicable to any biological molecule.

Current methods used to identify unknown proteins are typically similar to that illustrated in Fig. 12, but with the addition of database searching. The unknown protein is first cleaved into its constituent parts, as described above. The masses of the resulting constituent parts are analyzed and an experimental peptide map is generated. The determined masses are then compared against theoretical mass data generated for polypeptide sequences of a DNA (genome, cDNA, or otherwise) and/or protein database.

A biological molecule database is any compilation of information about

characteristics of biological molecules. Databases are the preferred method for storing both polypeptide amino acid sequences and the nucleic acid sequences that code for these polypeptides. The databases come in a variety of different types that have advantages and disadvantages when viewed as the hypothesis for a polypeptide identification experiment. The properties of the most common databases currently in use are listed in Table 4.

While the "database entry" for an amino acid sequence may appear to be a simple text file to a user browsing for a particular polypeptide, many databases are organized into very flexible, complicated structures. The detailed implementation of the database on a particular system may be based on a collection of simple text files (a "flat-file" database), a collection of tables (a "relational" database), or it may be organized around concepts that stem from the idea of a protein, gene, or organism (an "object-oriented" database).

Protein mass data may be predicted from nucleic acid sequence databases. Alternatively, protein mass data may be obtained directly from protein sequence databases which contain a collection of amino acid sequences represented by a string of single-letter or three-letter codes for the residues in a polypeptide, starting at the N-terminus of the sequence. These codes may contain nonstandard characters to indicate ambiguity at a particular site (such as "B" indicating that the residue may be "D" (aspartic acid) or "N" (asparagine). The sequences typically have a unique number-letter combination associated with them that is used internally by the database to identify the sequence, usually referred to as the accession number for the sequence.

Databases may contain a combination of amino acid sequences, comments, literature references, and notes on known posttranslational modifications to the sequence. A database that contains these elements is referred to as "annotated." Annotated databases are used if some functional or structural information is known

about the mature protein, as opposed to a sequence that is known only from the translation of a stretch of nucleic acid sequence. Non-annotated databases only contain the sequence, an accession number, and a descriptive title.

In general, each comparison of the unknown protein with the database proteins is assigned a score on the basis of a reasonable algorithm. Comparisons can be made and scores can be generated by a general purpose computer configured by software or otherwise. The unknown protein is then "identified" with a sequence that produces a score having a high degree of similarity.

More specifically, a score is a measure of the degree of similarity between the theoretical mass data of a database protein and the measured (experimental) mass data of an unknown protein for the same experimental conditions. The experimental conditions under which an unknown protein and the proteins from the database are handled should be the same.

Experimental conditions include the manner in which cleavage of the proteins is accomplished, that is, the specific substance used for the chemical degradation of the proteins. Additionally, the experimental condition defines the efficiency of the chemical degradation. The efficiency of a chemical degradation specifies the number of potential cleavage sites which may be expected to remain uncleaved. The mass data generated from the protein database may include mass data representing proteins with incomplete cleavages. Experimental conditions also include the method by which the mass data is generated.

Scores which denote a high degree of similarity are the top twenty scores generated in a comparison, more preferably the top ten scores, even more preferably the top five scores and most preferably the top one score.

A similarity between a group of measured masses of the unknown protein and a group of theoretical masses of a database protein is assessed by comparing every measured mass with every theoretical mass. A simple algorithm for the measure of similarity is the number of measured masses that are similar to at least one theoretical mass. For example, a measured peptide map of an enzymatically digested unknown protein can be compared with the theoretical masses calculated by applying the rules for the specificity of the enzyme to the amino acid sequence of a database protein.

More sophisticated algorithms can be used to generate a score. For example, ProFound (ProteoMetrics) is a software tool for searching protein sequence databases which measures similarity using a Bayesian statistical framework.

In the present invention a measured mass of an unknown protein and the theoretical masses of the proteins of the database are said to similar if the absolute value of the difference between them is less than the uncertainty in the measurement.

The similarity between the mass data of the unknown protein and the theoretical mass data of the database proteins is assessed taking into account the accuracy of the determination of the mass data by a particular method. For example, mass spectrometry determines a peptide mass m_i to an accuracy of $\pm\Delta m_i$, with $\Delta m_i/m_i$ typically >30ppm. Therefore, within the mass range $m_i \pm \Delta m_i$ peptide masses of several proteins in the database are considered to match the unknown protein.

The observed molecular mass or the observed isoelectric point of a protein can be used in combination with the measured masses of peptides generated by proteolysis to constrain the search for a polypeptide.

In particular, the comparison between the theoretical mass data of the database proteins and the mass data of the unknown protein may be constrained to only those

proteins of the database which are within a chosen mass range. The chosen mass range is preferably within 50% of the mass of the unknown protein, more preferably within 35%, most preferably within 25%.

Similarly, the comparison between the theoretical mass data of the database proteins and the mass data of the unknown protein may be constrained to only those proteins of the database which are within a chosen isoelectric point range. The isoelectric point (pI) of a protein is the pH at which its net charge is zero. The chosen isoelectric point range is preferably within 50% of the isoelectric point of the unknown protein, more preferably within 35%, most preferably within 25%.

Using the observed molecular mass or isoelectric point of a polypeptide to constrain a search must be done carefully. When nonannotated nucleotide sequence databases are used (such as TREMBL or GENPEPT), subsequent processing can greatly alter the pI or molecular mass of a protein, so much so that no identification can be made. For example, the small, highly conserved protein ubiquitin (SWISSPROT accession number P02248) has a molecular mass of 8.6 kD, which is the mass that would be measured by a mass spectrometer or a gel. A simple keyword search of the translated-nucleotide database GENPEPT results in several sequences for the same protein [accession numbers M26880 (77 kD), U49869 (25.8 kD) and X63237 (17.9 kD)]. None of these nucleotide-translated sequences give the correct molecular mass or pI, so using those parameters to limit a search would result in missing the database sequence altogether. Only annotated databases that fully outline known modifications can be used when the properties of the mature protein are being used to constrain a search.

Biological molecules may undergo common modifications in their structure. The mass data which is generated from a biological molecule database may include mass data representing biological molecules with common modifications.

Examples of such modifications are posttranslational modifications of proteins. The modification state of a protein is usually not known in detail. In database searches, it can be useful to assume that some common modifications might be present. This is achieved by comparing the measured peptides masses of the unknown protein with both the masses of the unmodified and modified peptides in the database.

Examples of posttranslational modifications include glycosylation and the oxidation of the amino acid methionine. Another example is the phosphorylation of the amino acids serine, threonine, and tyrosine. Phosphorylation is often used to activate or deactivate proteins and the phosphorylation state of an experimentally observed protein depends on may factors including the phase of the cell cycle and environmental factors.

Optionally, further information of the unknown protein's sequence is obtained by generating fragment mass data. Fragment mass data for a peptide can be generated in any manner which provides fragment mass data within a certain accuracy. Experimental conditions include the type of energy used to generate the fragment mass data. Vibrational excitation energy can be used. The vibrational excitation may be generated by collisions of the peptide with electrons, photons, gas molecules or a surface. Electronic excitation can be used. The electronic excitation may be generated by collisions of the peptide with electrons, photons, gas molecules (e.g. argon) or a surface.

In one embodiment, a measured fragment mass spectrum of a peptide from an enzymatically digested unknown protein is compared with the theoretical masses calculated by applying the rules for the specificity of the enzyme, and the rules for the fragmentation as known to those of ordinary skill in the art, to the amino acid

sequence of a database protein. For example, the software tool PepFrag (ProteoMetrics) allows for searching protein or nucleotide sequence databases using a combination of mass spectra data and fragmentation mass spectra data.

Fragment mass data for the purposes of this invention can be generated by using multidimensional mass spectrometry (MS/MS), also known as tandem mass spectrometry. A number of types of mass spectrometers can be used including a triple-quadruple mass spectrometer, a Fourier-transform cyclotron resonance mass spectrometer, a tandem time-of-flight mass spectrometer, and a quadruple ion trap mass spectrometer. Figure 13 illustrates this type of experiment. A single peptide from a protein digest is subjected to MS/MS measurement and the observed pattern of fragment ions is compared to the patterns of fragment ions predicted from database sequences.

All of the protein identification strategies outlined above to generate a score are currently available as CGI programs that can be accessed using a browser. Table 3 is a list of the programs available, as well as some of the relevant characteristics of the programs.

There is a risk of false identification of the unknown protein for several reasons. For example, each proteolytic peptide mass *measured* can be found in several proteins in a genome database. Also for example, a peptide map is often incomplete with respect to the protein identified and can contain a background of proteolytic peptide masses from other proteins. An identification of a protein is definitely uncertain if the result is characterized by a score that could as well be due to random matching between the peptide map and a protein in the database.

The present invention provides a method of assessing the statistical significance of a protein identification score for a particular experimental condition.

The method comprises generating a frequency distribution describing the null hypothesis (H_0). In the present specification, H_0 is defined as: "a protein identification is random and false."

A score frequency distribution for random protein identifications describes this null hypothesis. (For example see Figure 14.) A significance level (α) is chosen which represents the level of confidence desired for a particular protein identification. This significance level, along with the corresponding critical score (S_C), is indicated on the frequency distribution. A protein identification score is then compared to the frequency distribution. If the score falls within the significance level (α), the null hypothesis is rejected. The alternate hypothesis, "a protein identification is nonrandom," is then accepted. In other words, the null hypothesis should be rejected if the protein identification score is equal to or larger than the critical score since the critical score corresponds to the significance level.

The significance level of a protein identification result gives, in contrast with the identification score, an objective view of the quality of the result. However, significance testing can *never* tell if a result is true or false. Only biological methods have the potential of showing if a protein identification result is true.

The frequency of false results for repeated protein identifications using real data and significance testing at a level α is not generally predictable. The relative frequency of false results depends on the data as well as on the significance level chosen. A significant result is either false or true, as is a non-significant result. However, a general feature of significance testing is that if the significance level, α, is decreased, the relative frequency of false results considered to be significant is expected to decrease, and the relative frequency of true results considered non-significant is expected to increase. Optimized protein identification would, therefore, require (1) the use of an identification algorithm that maximizes the relative frequency

of true identifications, and (2) the use of significance testing at an appropriate significance level to discriminate against false identifications. Significance testing has the potential to reduce the relative frequency of false identifications independently of the identification algorithm used.

In one embodiment the invention provides a method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of proteins.

A frequency distribution is any compilation of the observed values of the variable being studied and how many times each value is observed. Frequency distributions can be in the form of a table of listings, a bar graph, a histogram, a frequency polygon, or a continuous curve. Functions derived from frequency distributions can be continuous (probability density function) or discrete (probability mass functions). Cumulative distribution functions of each type of function can also be derived.

The method comprises generating mass data for the particular experimental condition for known proteins from a protein sequence database as described above.

Mass data of a hypothetical protein for the same experimental condition is also generated. A hypothetical protein is any protein which is generated from selected peptides. These peptides may be selected from proteins which are selected from a database of proteins. These proteins and peptides may be selected by any method of selection. Selection may be accomplished by randomly selecting peptides from randomly selected proteins of a protein sequence database. Additionally, selection may be accomplished by selecting preselected peptides, such as the second peptides, of the selected proteins. Further selection may be limited to selecting the protein within a chosen mass range. The selected mass range may be from about 0.1 to about

3000kDa.

A sufficient number of peptides selected to generate a hypothetical protein may be in the range of about 1 to about 1000, more preferably from about 10 to 100.

Hypothetical proteins are preferably generated in such a manner that the set of peptides comprising a hypothetical protein is different from every set of peptides of the proteins in the database. In other words, the hypothetical protein does not appear in the protein database. Hypothetical proteins can be generated by a general purpose computer configured by software or otherwise.

Hypothetical proteins can also be generated by a mass spectrometer. A complicated mixture of proteins is enzymatically digested. Mass data is generated for these proteins.

The mass data, and optionally fragment mass data, generated for the hypothetical protein is compared with the data generated for each known protein in the database. The comparisons are carried out as described above. Since comparisons are made with a hypothetical protein, protein identifications are considered to be false and random. A score is calculated for each comparison. A score or scores are selected which correspond to the comparison denoting a high degree of similarity between the data. Additional and different hypothetical biological molecules are generated and comparisons performed until a sufficient quantity of scores are selected. A sufficient quantity of scores may, for example, be in the range of about 1 to 10^{10} scores, preferably 10 to 10^{8}, more preferably 50 to 10^{9} and most preferably from about 100 to about 10^{7}. The frequency of selecting each score is determined from which a score frequency distribution for random protein identifications is generated.

This invention provides a second method by which to generate a score

frequency distribution for random protein identifications. Mass data of a randomly selected protein are generated. In contrast to the hypothetical protein described above, which comprises peptides from different proteins, the randomly selected protein comprises a set of peptides from a single protein. The randomly selected protein is compared with database proteins and scores are calculated, as described above. The score which denotes a preselected degree of similarity, excluding the highest degree of similarity, between the randomly selected protein and a database protein is selected. An example is the score which denotes the second highest degree of similarity. Mass data of additional and different randomly selected proteins are generated and comparisons performed until a sufficient quantity of scores, as defined above, is selected. The frequency of selecting each score is determined from which a score frequency distribution for random protein identifications is generated.

As seen from Fig. 14, this second method yields a score frequency distribution which is very similar to the score frequency distribution generated by selecting the highest ranked protein identified on the basis of hypothetical proteins comprising maps with 20 tryptic peptides. If the size of the map is significantly greater than 20, only high-mass proteins could be the randomly selected proteins used in the second method (maximum number of tryptic peptides≈protein mass [Da]/1500 [Da]). These high-mass proteins have low abundance in the genome. The number of different maps available for large maps is therefore very limited, which would obscure the statistics in the score distribution for random matching.

In another embodiment the invention provides a method of determining the statistical significance of a protein identification score.

A significance level is selected which represents a level of confidence in a protein identification. The significance level is a function of a number of parameters, such as the number of masses in the peptide map, the mass accuracy, the degree of

incomplete enzymatic cleavage, the protein mass range, and the size of the genome. The Examples below illustrate the influence of these parameters on the score required for a fixed significance level.

The significance level may, for example, be any value in the range from about from 0.0001 to about 0.1, more preferably in the range from about 0.001 to about 0.05.

An identification score for an unknown protein is calculated wherein the score is associated with a known protein. The score is a function of similarity between mass data of the unknown protein and mass data generated for the database protein. The score is compared with a frequency distribution of random protein identification scores.

The frequency distribution indicates the selected significance level. It is determined whether the score associated with the unknown biological molecule identification is within the significance level. If the score is within the significance level the null hypothesis is rejected, and the identification is considered significant within the selected significance level.

In another embodiment the invention provides a method of identifying an unknown protein for a particular experimental condition and a particular significance level. The unknown protein can be in a mixture of proteins.

A significance level that would indicate a confident biological molecule identification is chosen; the unknown protein is cleaved into constituent parts under an experimental condition; mass data is generated for these constituent parts; the mass data is compared with mass data generated for the experimental condition from known proteins in a protein database; scores are calculated; and a score is selected, all as

described above. The selected score is compared with the frequency distribution described above. The frequency distribution indicates the chosen significance level along with the corresponding critical score. It is determined whether the selected score is equal to or larger than the critical score on the frequency distribution. If the score for the identification of the unknown protein is greater than or equal to the critical score then the null hypothesis, that the protein identification is random and false, is rejected.

If significant protein identification is not achieved directly by the methods described above, a researcher can try to obtain additional, further constraining information by tandem mass spectrometry that utilizes fragmentation of the proteolytic peptide ions in the mass spectrometer followed by analysis of the distribution of fragment ion masses and database searching. Results obtained from this technique should also be subjected to significance testing once a statistical basis has been founded by simulation.

Frequency functions and the score required for various significance levels for a variety of experimental constraints and for two different identification algorithms have been estimated. (See Examples.) With the critical score, S_C, known for a variety of experimental constraints, this invention provides significance testing which is fully automated and integrated with database searching software used for protein identification. This is a general method that will ultimately remove the difficulties associated with different algorithms yielding different scores.

It is to be appreciated that the methods or algorithms of the present invention described herein above may be performed using a general purpose computer or processing system which is capable of running application software programs, such as an IBM personal computer (PC) or suitable equivalent thereof. Preferably, the application program code is embedded in a computer readable medium, such as a

floppy disk or computer compact disk (CD). Furthermore, the computer readable medium may be in the form of a hard disk or memory (e.g., random access memory or read only memory) included in the general purpose computer.

As appreciated by one skilled in the art, the computer software code may be written, using any suitable programming language, for example, C or Pascal, to configure the computer to perform the methods of the present invention. While it is preferred that a computer program be used to accomplish any of the methods of the present invention, it is similarly contemplated that the computer may be utilized to perform only a certain specific step or task in an overall method, as determined by the user.

Preferably, the methods of the present invention are used with one or more displays (e.g., conventional CRT or liquid crystal display) provided with the processing system for presenting an indication of, for example, the final result of the process or algorithm. The display may preferably be utilized to present such information graphically (e.g., charts or three dimensional models of biological molecules) for further clarity.

In addition to performing the necessary calculations and processing functions in accordance with the present invention, the general purpose computer may also be used, for example, to store data pertaining to known biological molecules corresponding to a predetermined experimental condition. Such information may be stored on a hard disk or other memory, either volatile or non-volatile, included in the computer. Similarly, the information may be stored on a computer readable medium, such as floppy disk or CD, which can be transported for use on another computer system, as appreciated by those skilled in the art. In this manner, the methods of the present invention may be performed on any suitable general purpose computer and are not limited to a dedicated system.

Those of ordinary skill in the art will recognize that the present invention has wide applicability for identification of unknown biological molecules. Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be effected therein by one skilled in the art without departing from the scope or spirit of the present invention.

EXAMPLES

Basic Protein Identification

Figure 3a shows a MALDI ion trap mass spectrometry (ITMS) spectrum of a tryptic digest of an unknown protein form *Saccharomyces cerevisiae* that was observed as a spot on a gel. The spectrum has more than a dozen major peaks. If the corresponding masses are used to search all *S. cerevisiae* sequences in OWL with ProFound, a list of proteins that are most likely to give the observed tryptic map is obtained (Table 2). In this example, submit P130 of eukaryotic initiation factor 4F (IF42__YEAST) is the most probable protein. To further increase the confidence in the identification, the ions with m/z=2596 were isolated and fragmented (Fig. 3b). The spectrum contains three major fragment ions. The peak at m/z=2468 is loss of the C-terminal lysine and contains little information. The two other fragment peaks, on the other hand, correspond to fragmentation at the C-terminal side of acidic amino acids. If a database is searched for proteins that have tryptic peptides with mass 2595 Da that fragment at the C-terminal side of acidic amino acids to give rise to b or y ions with mass 1984 and 2337 Da there is only one yeast protein (IF42__YEAST) in the public databases that agree with this information. The tryptic peptide is AQPISDIYEFAYPENVERPDIK and the two fragment ions correspond to fragmentation on the C-terminal side of the aspartic acids at residues 6 and 20, respectively. If a theoretical trypsin digest of IF42__YEAST is compared with the peptide map, all major peaks can be assigned to tryptic peptides from IF42__YEAST or to peptides from autolysis of trypsin.

Generation of a Hypothetical Protein and Significance Testing

MATERIALS AND METHODS

The method designed to estimate frequency functions for random protein

identification involves two steps: (1) generation of random peptide maps from a genome and (2) simulation of protein identification by searching a genome database and using the random peptide maps as data.

In the following experiments the proteins were digested with trypsin. However, the significance testing method can be generalized to any protein identification experiment. The simulations can be generalized to all reasonable alternatives of proteases as well as tandem mass spectrometry.

Tryptic Peptide Maps

Random tryptic peptide maps (trypsin cleaves with high specificity at the carboxyl side of lysine and arginine) were generated from tryptic peptide masses predicted by the open reading frames (ORF:s) of a genome database. Each tryptic peptide mass was randomly chosen from a single randomly chosen protein in the database. Each protein was not allowed to contribute more than one peptide mass to a particular map. Typically only completely digested tryptic peptides were used, but also maps with various degrees of incomplete cleavage in the tryptic peptides were used for comparison. Various *data sets* of random tryptic peptide maps were generated. Within each data set, all maps contained the same number (in the range 6-80) of tryptic peptide masses.

The genome databases used were: *Haemophilus influenzae*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, containing 1,718, 6,403, and 16,332 ORF:s respectively. *S. cerevisiae* was used predominantly and the other two databases only to study the influence of the size of the genome on the score distribution for random protein identification.

Simulation of Protein Identification due to Random Matching

Each map of a data set was subjected to protein identification by database searching. Two different identification algorithms were employed that will be referred to as algorithm 1 and algorithm 2. Algorithm 1 is a streamlined version the ProFound algorithm (publicly available through the World Wide Web, http://prowl), which ranks the proteins according to a Bayesian likelihood calculated as the peptide map is compared with the database proteins. Algorithm 1 takes into account the number of matches between a database protein and the peptide map within the accuracy of the mass measurement, but also weighs in indirectly the protein mass as well as the expected efficiency of the protease used in an experiment. Algorithm 2 ranks proteins only according to their number of matches with tryptic peptide masses in the peptide map. In the simulations, the score and the name of the highest ranked protein as well as the source protein of each random tryptic peptide mass were stored for each random tryptic peptide map of a data set. Random and true identifications and random and false identifications could be distinguished. If more than one protein was identified and their sequences were not similar, the result was interpreted as false. A simulation with a set of different random peptide maps of the same size yields a distribution of the score for random protein identifications characteristic for that peptide map size and other constraints used in the database search. The typical parameters used in the simulations are summarized in Table 1. However, the experimentally pertinent parameters were varied systematically, one by one, in order to measure their respective influence on the identification score distribution.

The code for generating peptide maps as well as for the simulation of protein identification was written in C. A script written in Perl was employed for processing of the simulation results. All simulations were performed on a Dell XPS (300 MHZ Pentium Pro) personal computer.

Significance Testing

The simulated frequency distributions of the protein identification score for a set of random tryptic peptide maps of size 20 and 80 respectively are shown in Fig. 1. Apparently the fraction of the map that matches a database protein by chance drops as the map size increases (Fig. 1b). Knowledge of score frequency distributions of the type shown in Fig. 1 is the basis for testing the significance of protein identifications.

In the simplest form of significance testing, a null hypothesis H_0 is either rejected or not rejected at some *significance level* α. Here, H_0 is defined as: *"a protein-identification is random and false."* In order to test the significance of an experimental result, the frequency function $f(S)$ describing H_0 must be known or estimated. By simulating many protein identifications using *random* tryptic peptide maps and by storing each identification score S while distinguishing between random and true (rare events) and random and false results, $f(S)$ could be estimated simply by dividing a score frequency distribution of the type shown in Fig. 1 by the number of simulated protein identifications (i.e. peptide maps). The problem of testing if a protein identification result deviates significantly from H_0 falls naturally into the category of one-sided significance testing using the protein identification score resulting from the experiment as the test variable. That is, if the score S is larger than a critical value S_C, H_0 is rejected, otherwise it is not rejected. S_C is derived from the equation:

$$\sum_{S \geq S_C} f(S) \leq \alpha$$

(for a discrete distribution), where α is a significance level (sometimes called *test error risk*) chosen prior to the significance testing. α represents the statistical risk (probability) that H_0 would be rejected if it actually were true. Apparently α should be

small, and often the values 0.05, 0.01, or 0.001 are chosen. Fig. 2 illustrates the entire procedure of employing simulated score distributions to estimate frequency functions and to find what scores correspond to a particular significance level.

The Number of Masses in the Peptide Map

In order to make significance testing applicable; one has to find the score S_C required for a significance level α under the particular conditions of an experiment. For example, the number of tryptic peptide mass peaks can vary considerably between different experiments. The influence of the number of peaks on the score required for significant identification is illustrated in Fig. 3. It is seen clearly that a large peptide map is desirable in order to discriminate against false identifications.

Mass Accuracy

The mass accuracy Δm in an experiment is typically entered in the database search. The entry should reflect the true Δm. State of the art instruments typically employed for peptide mapping can provide $\Delta m \leq 0.1$ Da for peptides. Here, Δm was varied between 0.006 and 1 Da. It is seen from Fig. 4, where S_C is plotted versus Δm, that a further improvement of Δm from the presently typical level with for example one order of magnitude would clearly facilitate unambiguous protein identifications. The local minima of the functions fitted to the data simulated with algorithm 1 (Fig. 4 top panel), are due to the non-uniform distribution of peptide masses in the database and that the score is proportional to $(\Delta m)^{-r}$, where r is the number of matches. Algorithm 2 yields a plateau due to the non-uniform distribution of peptide masses (Fig. 4, bottom panel).

Incomplete Enzymatic Cleavage

Enzymatic digestion is often imperfect. The expected highest number u of

specific sites not cleaved in an experiment is typically entered as a constraint in the database search. Here, u was varied between 0 and 4 in different database searches. It is seen from the results in Fig. 5, where S_C is plotted as a function of u, that an as complete cleavage as possible reduces the number of possibilities for random matching of peptide masses and therefore facilitates the discrimination against false results.

Algorithm 1 has an intrinsic moderation of the influence on the score S by the number N of possible proteolytic peptide masses per database protein: S is proportional to $(N-r)!/N!$, where r is the number of matches. This moderation causes a drop of S_C for large values of u, as the number of matches either saturates (Fig. 5a) or begins to increase less steeply with u (Fig. 5b).

Independently of algorithm, the trend of saturation shown in Fig. 5 illustrates that poor cleavage does not necessarily ruin an experiment, and incomplete cleavage can in fact sometimes yield further information. This extra information is not utilized in any of the scoring algorithms applied here.

Protein Mass

Due to potentially occurring protein degradation and anomalous migration of modified proteins in SDS-gel electrophoresis, mass information from gels as a constraint for protein identification should be used with caution. In most of the simulations, the protein mass was restricted to 100 kDa for the proteins generating the peptide maps as well as in the database search. About 95% of *S. cerevisiae* proteins are within this mass range. In order to cover the remaining 5% of proteins, the influence on the score required for significance as a function of the protein mass range was studied. Fig. 6 shows that algorithm 1 is rather insensitive to an increased protein mass range, whereas algorithm 2 yields a high degree of random matching with high-

mass proteins.

Genome Size

The score required for significance, S_C, was studied as a function of the size of the genome. The results shown in Fig.7 are based on data (random tryptic peptide maps) from a prokaryote *H. influenzae*, a single-cell eukaryote *S. cerevisiae* (budding yeast), and a multicellular organism *C. elegans* (nematode) respectively. The yeast genome was divided into two parts of similar size and the results from the respective parts were averaged. The score required for significant protein identification increases with the size of the genome (Fig.7). Identification of *H. influenzae* and *C. elegans* proteins from random tryptic peptide maps generated from the *S. cerevisiae* genome yielded similar results as those of Fig. 7 (data not shown). This is due to the highly similar distribution of tryptic peptide masses for different genomes. Therefore, the dependence of S_C on the genome size shown here can be used to estimate S_C for any genome within the size range studied.

Statistical Uncertainties

The number of protein identifications simulated and the shape of the score frequency function can influence the accuracy of S_C, the score required for a significant result. How sensitive the value of S_C is to statistical fluctuations was probed carefully by (1) repeated simulation with each simulation performed under identical conditions except for the set of random numbers used to generate the maps, and (2) by varying the number of random peptide maps used per simulation. The pronounced discrete nature of the score distribution of algorithm 2 implies an inherent sensitivity to statistical fluctuations in the simulations. When using algorithm 2 and maps composed of <35 peptides, the values of S_C for α=0.01 and α=0.001 did not converge (jumped e.g. between 8 and 9 matches) as the number of peptide maps per

simulation was increased (≤15000). For these significance levels and algorithm 2, the highest S_C observed was assumed as the result. For larger peptide maps this potential uncertainty is reduced due to a broadening of the score frequency distribution (Fig. 1b). Significance testing with heavily discrete distributions can also cause a fairly large discrepancy between the values of α and P, where P is the area under the score frequency function $f(S)$ for $S \geq S_C$ (Fig. 2). For example, in the results for α=0.05 plotted in Fig. 3 the median of the relative deviation $(\alpha-P)/\alpha$ was 0.63 for algorithm 2, but only 0.12 for algorithm 1.

The less discrete nature of algorithm 1 allows even minor statistical fluctuations to be resolved and examined. As several simulations with peptide maps generated by different random numbers were performed under identical conditions, the fluctuation of S_C due to a different response to different random data could be probed. The standard deviation of the mean S_C derived from five different simulations decreased sharply when going from 500 to 1000 maps per simulation and then changed very moderately when the number of maps was increased further. For five simulations each using 1000 maps with 20 random tryptic peptide masses, the relative standard deviation from the mean S_C was 0.6%, 1.5%, and 2.5% for the 0.05, 0.01, and 0.001 significance levels respectively. The mean S_C determined with 1000 maps per simulation differed by 0.4% from the S_C determined by simulating with 15,000 maps. Hence, for algorithm 1 the magnitude of S_C is well established already at 1000 maps per simulation, but statistical fluctuations between different data sets can in an individual simulation cause an uncertainty of a few percent.

To avoid potential bias by statistical fluctuations between different sets of random peptide maps (different simulations), the same set of random peptide maps was used when the respective functions of mass accuracy, the number of uncleaved sites, and maximum protein mass were derived (Figs. 4-6). For the function describing the dependence on the maximum number of uncleaved sites (Fig. 5), the

concept of using the same data-set was utilized by using random maps with completely cleaved peptides while varying the maximum number of uncleaved sites *u* in the database search. However, the relation between the relative score and *u* was the same as that plotted in Fig. 5 if instead the same *u* was used for the tryptic peptides in the random maps as in the database search (data not shown). For the function describing the dependence on the maximum protein mass (Fig. 6), random peptide maps generated with no limitation on the mass of the proteins contributing to the maps were used, and the protein mass range was varied in the database search only. A comparison was made between the results plotted in Fig. 6 (case 1) and results from simulations using the same protein mass limitations in the peptide maps as in the database search (case 2). For random maps with 20 tryptic peptides no difference was observed between case 1 and case 2. For simulations with 50 peptides in the random tryptic peptide maps, algorithm 1 displayed no clear difference between the two cases, whereas algorithm 2 yielded a slightly less pronounced dependence on the protein mass for case 2. The latter observation is due to that the random and true results, which occur at a frequency increasing with the number of peptides in the maps, are somewhat suppressed when the limitation of the protein mass is in the database search only. For example, if a protein with mass 400 kDa contributes with a peptide to a map, that protein can never be identified by chance when the database search is done with a protein mass limit of 300 kDa.

The approach of fitting score frequency functions to appropriate analytical functions could be a means of reducing difficulties associated with discrete distributions and statistical fluctuations. Such an approach remains to be explored for simulation of random protein identification, but has been employed for determination of scores required for significance in sequence or structure comparison algorithms.

Exploring a Large Parameter Space

The results presented in Figs. 1-7 are based on 110 simulations including the two different scoring methods (>10^5 protein identifications). Clearly, these simulations represent a small fraction even of the limited range of various search parameters studied. To generate a mesh within which interpolation could be accurately performed would require about 10^7 protein identifications per algorithm. This is quite feasible and would require about 30 days of computation time per algorithm on a regular PC. A less computationally intensive approach is to use the functions that describe how S_C varies in the parameter space studied (Figs. 3-7). We tested this approach by comparing such estimations of S_C with values of S_C derived from direct simulations for randomly chosen points in the parameter space. A very good correlation between estimation and simulation was obtained (0-5% deviation, similar to the uncertainty in S_C from a single simulation). Therefore, estimation based on the functions derived here is an accurate procedure to assess S_C in an arbitrary point in the parameter space.

The Design of Tryptic Peptide Maps

The hypothetical data used in the protein identification simulations were random tryptic peptide maps. These maps were used with the specific goal to elucidate the distribution of scores due to random matching between the masses of the hypothetical tryptic peptide map and theoretical masses of tryptic peptides of proteins in the database. The random matching can be studied in various ways. Completely random peptide maps, where each proteolytic peptide mass is generated from a different protein, and hence are non-correlated, is used here. The random peptide maps represent an extreme protein mixture (or one or more heavily modified proteins), that might be unrealistic. The null hypothesis H_0: *the protein identification is false and random*, could be investigated by data that are closer to what a researcher

would expect to observe experimentally. But, the significance testing of the deviation from H_0 does not suggest that real data would often strongly resemble the score distribution for random matching. The interesting part of the random score distribution is the high-score-side where the score distribution of real data could presumably overlap, hence making the real and the random distributions indistinguishable. If simulation with data from 6 proteins in a map with 20 tryptic peptides is done, and all possible combinations of the number of peptides per protein in the map with equal abundance are used, the number of true results increases compared with the completely random maps, and the true results typically yield higher scores. Apparently, true and false results can also yield similar scores. As expected, the scores of the false results are similar if the map is constructed from 6 or from 20 proteins. However, the small number of false results yielded with 6 proteins provide poor statistics for the score distribution that we need to describe H_0, and hence the estimate of the score corresponding to significance would become considerably more uncertain.

Genome	S. CEREVISIAE
Number of tryptic peptides in a map	N
Number of uncleaved sites in map peptides	0
Number of proteins contributing to a map	N
Number of maps used in a simulation	1000
Maximum number of uncleaved sites allowed in the database search	2
Mass accuracy in the database search [Da]	0.1
Maximum protein mass in data generation and database searches [kDa]	100

Table 1. Simulation Parameters

The protein identification simulation conditions employed unless otherwise stated.

Table 2 Result from a ProFound Search of All *S. Cerevisiae* Proteins in OWL.

1.	IF42_YEAST (probability = 7.9e-01) EUKARYOTIC INITIATION FACTOR 4F SUBUNIT P130
2.	NPR1_YEAST (probability = 2.8e-02) NITROGEN PERMEASE REACTIVATOR PROTEIN
3.	S63138 (probability = 1.9e-02) PROBABLE PROTEIN KINASE NPR1
4.	RL3E_YEAST (probability = 1.8e-02) 60S RIBOSOMAL PROTEIN L30E
5.	SCYOR206W (probability = 1.5e-02) HYPOTHETICAL PROTEIN SCYOR206W
6.	CBF5_YEAST (probability = 1.4e-02) CENTROMERE/MICROTUBULE BINDING PROTEIN CBF5
7.	S52893 (probability = 9.5e -03) HYPOTHETICAL PROTEIN YMR044W

Note: m/z = 2164 and 2275 were not used in the search, because they are from autolysis of trypsin.

Table 3 CGI Programs Currently Used for Protein Identification

CGI program	Data type	Databases	Cleavage chemistry	Additional search parameters	Help	AI	Report format
PepFrag [21]	MS/MS	10 (aa + na)	7 Enzymes + CNBr	Protein mass, peptide masses, taxonomy, missed cleavage sites, daughter ion types, mass accuracy	Pages + examples	No	Active + tools
ProFound [22]	Masses	1 (aa)	5 Enzymes + CNBr	Protein mass, taxonomy, missed cleavage sites, mass accuracy	Pages + examples	Yes	Active + tools
MS-Fit [23]	Masses	5 (aa + na)	9 Enzymes + CNBr	Protein mass, taxonomy, missed cleavage sites, mass accuracy	1 Example	No	Active
MS-Tag [24]	MS/MS	5 (aa + na)	9 Enzymes + CNBr	Protein mass, taxonomy, missed cleavage sites, daughter ion types, mass accuracy	1 Page	No	Active
PeptideSearch [25]	Masses	1 (na)	7 Enzymes + CNBr	Protein mass, mass accuracy	None	No	Active
PeptideSearch [26]	MS/MS	1 (na)	7 Enzymes + CNBr	Protein mass, daughter ion types, mass accuracy	None	No	Active
Mass Search [27]	Masses	2 (aa + na)	11 Enzymes + 7 chemical	Protein mass	1 example	No	Dead
Mowse [28]	Masses	1 (aa)	6 Enzymes + CNBr	Protein mass, amino acid composition, mass accuracy	1 Page	Yes	Active

Note: Abbreviations: aa, amino acid sequence database; na, nucleic acid sequence database; CNBr, cyanogen bromide; and AI, any type of artificial intelligence/ expert system used to rate results. An "active" report format has links that you can follow to get the sequence of a matched protein, and "tools" refer to additional programs to help make use of that sequence information. A "dead" report just lists the matching proteins' names and the user is expected to find the sequence information and pass judgement on the results without any assistance.

SUBSTITUTE SHEET (RULE 26)

Table 4 Sequence Databases Currently Used for Protein Bioinformatics

Database	Peptide sequences	Nucleic acid sequences	Peptide annotation	Redundancy	Sequence reliability	Entries (/1000)
SWISSPROT [3]	Yes	No	Complete	Low	Excellent	59
PIR [4]	Yes	No	Complete	Moderate	Excellent	95
EMBL [5]	No	Yes	Some	High	Good	1438
TREMBL [5]	Yes	No	Some	High	Good	137
GENBANK [6]	No	Yes	Some	High	Good	1766
GENPEPT[6]	Yes	No	Some	High	Good	262
OWL [7]	Yes	No	Some	Low	Good	210
dbEST [8]	No	Yes	None	Very High	Low	1317

Note: The sequence reliability for GENBANK and EMBL refers to the normal nucleic acid sequences stored in those databases. These databases also include expressed sequence tag data, which are not as reliable.

SUBSTITUTE SHEET (RULE 26)

WE CLAIM:

1. A method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules, the method comprising:

a) generating mass data for the particular experimental condition for known biological molecules in a biological molecule database;

b) generating mass data of a hypothetical biological molecule for the experimental condition;

c) comparing the data generated in step (b) with the data generated for each known biological molecule in step (a);

d) calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data generated in step (a), which corresponds to a particular known biological molecule, and the data generated in step (b);

e) selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the data generated in step (a) and the data generated in step (b);

f) repeating steps (b) through (e) with different hypothetical biological molecules until a sufficient quantity of scores are selected; and

g) determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

2. The method according to Claim 1 wherein the frequency distribution comprises a function derived from the frequency distribution.

3. The method according to Claim 2 wherein the function derived from the frequency distribution is a discrete function.

4. The method according to Claim 2 wherein the function derived from the frequency distribution is a probability density function.

5. The method according to Claim 1 wherein the mass data in step (a) are generated by a computer.

6. The method according to Claim 1 wherein the mass data in step (b) are generated by a computer.

7. The method according to Claim 1 wherein the mass data in step (b) are generated by a mass spectrometer.

8. The method of Claim 1 wherein the known biological molecules and hypothetical biological molecules are proteins.

9. The method of Claim 1 wherein the known biological molecules and hypothetical biological molecules are nucleic acid molecules.

10. The method of Claim 1 wherein the known biological molecules and hypothetical biological molecules are polysaccharides.

11. The method according to Claim 1 wherein a sufficient quantity of scores is in the range of from about 1 to 1010 scores.

12. The method according to Claim 1 wherein a sufficient quantity of scores is in the range of from about 100 to 107 scores.

13. The method according to Claim 1 wherein the experimental condition

defines the mass data as resulting from chemical degradation of the known biological molecules and hypothetical biological molecules.

14. The method according to Claim 13 wherein the chemical degradation is enzymatic digestion.

15. The method according to Claim 13 wherein the experimental condition defines an efficiency of the chemical degradation.

16. The method of Claim 14 wherein the enzymatic digestion is by trypsin.

17. The method according to Claim 1 wherein the comparison in step (c) is constrained to known biological molecules within a chosen mass range.

18. The method according to Claim 1 wherein the comparison in step (c) is constrained to known biological molecules within a chosen isoelectric point range.

19. The method according to Claim 1 wherein the experimental condition defines a particular accuracy for mass data determination.

20. The method according to Claim 1 wherein the comparison in step (c) comprises known biological molecules which exhibit modifications.

21. The method according to Claim 20 wherein the modifications of the biological molecules are posttranslational modifications of proteins.

22. The method according to Claim 1 wherein the hypothetical biological molecule is generated by a method which comprises:

a) selecting at least one known biological molecule from a biological

molecule database;

b) generating mass data of the known biological molecule in step (a);

c) selecting a mass from step (b) which corresponds to at least one constituent part of the known biological molecule;

d) repeating steps (a) through (c) for a different selected biological molecule until a sufficient number of masses are selected to generate a hypothetical biological molecule.

23. The method according to Claim 22 wherein the biological molecule in step (a) is randomly selected.

24. The method according to Claim 22 wherein the constituent part of the biological molecule in step (c) is randomly selected.

25. The method according to Claim 22 wherein the hypothetical biological molecule comprises a set of constituent parts which is different from every set of constituent parts of the known biological molecules of the biological molecule database.

26. The method of Claim 22 wherein the constituent parts comprises peptides.

27. The method of Claim 22 wherein the constituent parts comprises oligonucleotides.

28. The method of Claim 22 wherein the constituent parts comprises polysaccharides.

29. The method according to Claim 22 wherein a sufficient number of

masses selected in step (c) is in the range of from about 1 to about 1000.

30. The method according to Claim 22 wherein the biological molecule of step (a) is within a chosen mass range.

31. The method according to Claim 30 wherein the chosen mass range is from about 0.1 to about 3000 kDa.

32. The method according to Claim 1 wherein fragment mass data is generated for at least one constituent part of the known biological molecules and hypothetical biological molecule.

33. The method according to Claim 32 wherein the comparison between data for each of the known biological molecules and the hypothetical biological molecule comprises the comparison of the fragment mass data.

34. The method according to Claim 32 wherein the experimental condition defines the energy used to generate the fragment mass data.

35. A method of generating a frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules, the method comprising:

a) generating mass data for the particular experimental condition for known biological molecules in a biological molecule database;

b) randomly selecting a biological molecule from the database

c) comparing the mass data of the randomly selected biological molecule of step (b) with the mass data of each known biological molecule in step (a);

d) calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data of step (a), which corresponds to a

particular known biological molecule, and the data of the randomly selected biological molecule of step (b);

e) selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a degree of similarity between the data generated in step (a) and the data of the randomly selected biological molecule of step (b) which is lower than the highest degree of similarity;

f) repeating steps (b) through (e) with different randomly selected biological molecules until a sufficient quantity of scores are selected; and

g) determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

36. A method of identifying an unknown biological molecule for a particular experimental condition and a particular significance level, the method comprising:

a) selecting a significance level that represents a level of confidence in a biological molecule identification;

b) cleaving the unknown biological molecule into constituent parts;

c) generating mass data for these constituent parts;

d) comparing the mass data generated in step (c) with mass data generated for the experimental condition from known biological molecules of a biological molecule database;

e) calculating scores for each comparison in step (d), wherein the scores are a function of similarity between mass data of the unknown biological molecule and mass data generated from the biological molecule database;

f) selecting a score generated in step (e) wherein the score corresponds to a comparison which denotes a high degree of similarity and wherein the score corresponds to a particular known biological molecule in the biological molecule database;

g) comparing the score selected in step (f) with a frequency distribution

of scores for the experimental condition, wherein the distribution relates to random biological molecule identifications, and wherein the frequency distribution has associated therewith a critical score which corresponds to the significance level;

and

h) determining whether the score selected in step (f) is equal to or larger than the critical score.

37. The method according to Claim 36 wherein the unknown biological molecule is cleaved into constituent parts by a method that produces constituent parts in a predictable way.

38. The method according to Claim 36 wherein the mass data of step (c) are generated by a mass spectrometer.

39. The method according to Claim 36 wherein the unknown biological molecule is in a mixture of biological molecules.

40. The method according to Claim 36 wherein the comparison in step (d) is constrained to known biological molecules within a chosen mass range.

41. The method according to Claim 40 wherein the chosen mass range is within 25% of the mass of the unknown biological molecule.

42. The method according to Claim 40 wherein the chosen mass range within is from about 0.1 to about 3000 kDa.

43. The method according to Claim 36 wherein the experimental condition defines the mass data as resulting from chemical degradation of the known biological molecules and hypothetical biological molecules.

44. The method according to Claim 43 wherein the chemical degradation is enzymatic digestion.

45. The method according to Claim 43 wherein the experimental condition defines efficiency of the chemical degradation.

46. The method according to Claim 36 wherein the comparison in step (d) is constrained to biological molecules of the database which have an isoelectric point within a particular range.

47. The method according to Claim 46 wherein the isoelectric point range is within 25% of the isoelectric point of the unknown biological molecule.

48. The method according to Claim 36 wherein the experimental condition defines a particular accuracy for the determination of the mass data.

49. The method according to Claim 36 wherein the comparison in step (d) comprises known biological molecules which exhibit modifications.

50. The method according to Claim 36 wherein fragment mass data is generated for at least one constituent part of the known biological molecules and hypothetical biological molecules.

51. The method according to Claim 50 wherein the comparison between data for each of the known biological molecules and the hypothetical biological molecule comprises the comparison of the fragment mass data.

52. The method according to Claim 50 wherein the experimental condition

defines the energy used to generate the fragment mass data.

53. The method according to Claim 52 wherein the fragment mass data is generated by vibrational excitation.

54. The method according to Claim 52 wherein the fragment mass data is generated by electronic excitation.

55. The method according to Claim 53 wherein the vibrational excitation is generated by collisions with electrons, photons, gas molecules or a surface.

56. The method according to Claim 54 wherein the electronic excitation is generated by collisions with electrons, photons, gas molecules or a surface.

57. A method of determining statistical significance of a biological molecule identification score, the method comprising:

a) selecting a significance level that represents a level of confidence in a biological molecule identification;

b) calculating a score associated with an unknown biological molecule, wherein the score is a function of similarity between mass data of the unknown biological molecule and mass data generated for known biological molecules of a biological molecule database;

c) comparing the score with a score frequency distribution, wherein the distribution is generated by comparing mass data of hypothetical biological molecules with mass data generated for known biological molecules of a biological molecule database, and wherein the frequency distribution has associated therewith the significance level; and

d) determining whether the score associated with the unknown biological molecule identification is within the significance level.

58. The method according to Claim 57 wherein the significance level denotes probability that a biological molecule identification is false.

59. A frequency distribution of scores for a particular experimental condition, wherein the scores relate to random identifications of biological molecules, the method comprising:

a) a means for generating mass data for the particular experimental condition for known biological molecules in a biological molecule database;

b) a means for generating mass data of a hypothetical biological molecule for the experimental condition;

c) a means for comparing the data generated in step (b) with the data generated for each known biological molecule in step (a);

d) a means for calculating a score for each comparison in step (c), wherein the score is a function of similarity between the data generated in step (a) which corresponds to a particular known biological molecule and the data generated in step (b);

e) a means for selecting a score from the scores calculated in step (d), wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the data generated in step (a) and the data generated in step (b);

f) a means for repeating steps (b) through (e) with different hypothetical biological molecules until a sufficient quantity of scores are selected; and

g) a means for determining the frequency of selecting each score and generating therefrom a frequency distribution of scores.

60. A means for identifying a biological molecule for a particular experimental condition and a particular significance level comprising:

a) a means for selecting a significance level that represents a level of confidence in a biological molecule identification

b) a means for cleaving the unknown biological molecule into constituent parts by a method that produces constituent parts;

c) a means generating mass data for these constituent parts;

d) a means comparing the mass data generated in step (c) with mass data generated for the experimental condition from known biological molecules of a biological molecule database;

e) a means calculating scores for each comparison in step (d), wherein the scores are a function of similarity between mass data of the unknown biological molecule and mass data generated from the biological molecule database;

f) a means selecting a score generated in step (e) wherein the score corresponds to a comparison which denotes a high degree of similarity and wherein the score corresponds to a particular known biological molecule in the biological molecule database; and

g) a means determining whether the score selected in step (f) is equal to or larger than the critical score.

61. A means for determining statistical significance of a biological molecule identification score, the method comprising:

a) a means for selecting a significance level that represents a level confidence in a biological molecule identification;

b) a means for calculating a score associated with the unknown biological molecule, wherein the score is a function of similarity between mass data of the unknown biological molecule and mass data generated from a biological molecule database;

c) a means for comparing the score with a score frequency distribution, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated from a biological molecule database, and wherein the frequency distribution has associated therewith a significance level determined to represent a confident biological molecule identification; and

d) a means for determining whether the score associated with the unknown biological molecule identification is within the significance level.

62. A computer program product comprising:

a computer usable medium having computer readable program code means embodied in said medium for generating a frequency distribution of scores, wherein the scores relate to random identifications of biological molecules, said computer program product including:

computer readable program code means for causing a computer to generate mass data for each known biological molecule in a biological molecule database for a particular experimental condition;

computer readable program code means for causing the computer to generate mass data of a hypothetical biological molecule for the experimental condition;

computer readable program code means for causing the computer to compare the mass data of the hypothetical biological molecule with the mass data generated for each known biological molecule in the biological molecule database for the particular experimental condition;

computer readable program code means for causing the computer to calculate a score for each mass data comparison, wherein the score is a function of similarity between the mass data corresponding to a particular known biological molecule and the mass data corresponding to the hypothetical biological molecule;

computer readable program code means for causing the computer to select a score from the calculated scores, wherein the selected score corresponds to the comparison which denotes a high degree of similarity between the mass data corresponding to the particular known biological molecule and the mass data corresponding to the hypothetical biological molecule;

computer readable program code means for causing the computer to repeatedly generate mass data of different hypothetical biological molecules, compare

the mass data each of the hypothetical molecules with the mass data generated for each known biological molecule in the biological molecule database, calculate a score for each of the mass data comparisons and select a score from the calculated scores until a sufficient quantity of scores are selected; and

computer readable program code means for causing the computer to determine the frequency of selecting each score and to generate therefrom a frequency distribution of scores.

63. A computer program product comprising:

a computer usable medium having computer readable program code means embodied in said medium for identifying an unknown biological molecule for a particular experimental condition and a particular significance level, said computer program product including:

computer readable program code means for causing a computer to generate mass data of an unknown biological molecule, the unknown biological molecule having been cleaved into constituent parts by a method that produces constituent parts;

computer readable program code means for causing the computer to compare the mass data of the unknown biological molecule with mass data generated for the experimental condition from known biological molecules of a biological molecule database;

computer readable program code means for causing the computer to calculate scores for each mass data comparison, wherein the scores are a function of similarity between mass data of the unknown biological molecule and mass data generated from the biological molecule database;

computer readable program code means for causing the computer to select a score from the calculated scores, wherein the selected score corresponds to a particular known biological molecule in the biological molecule database, and wherein the selected score corresponds to a comparison which denotes a high degree

of similarity;

computer readable program code means for causing the computer to compare the selected score with a frequency distribution of scores for the experimental condition, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated from a biological molecule database, and wherein the frequency distribution has associated therewith a critical score which corresponds to the significance level; and

computer readable program code means for causing the computer to determine whether the selected score is equal to or larger than the critical score.

64. A computer program product comprising:

a computer usable medium having computer readable program code means embodied in said medium for determining statistical significance of a biological molecule identification score, said computer program product including:

computer readable program code means for causing a computer to calculate a score associated with an unknown biological molecule, wherein the score is a function of similarity between mass data of the unknown biological molecule and mass data generated from a biological molecule database;

computer readable program code means for causing the computer to compare the score with a score frequency distribution, wherein the distribution is generated by comparing mass data of a hypothetical biological molecule with mass data generated from a biological molecule database, and wherein the frequency distribution has associated therewith a significance level determined to represent a confident biological molecule identification; and

computer readable program code means for causing the computer to determine whether the score associated with the unknown biological molecule identification is within the significance level.

FIG. 1a

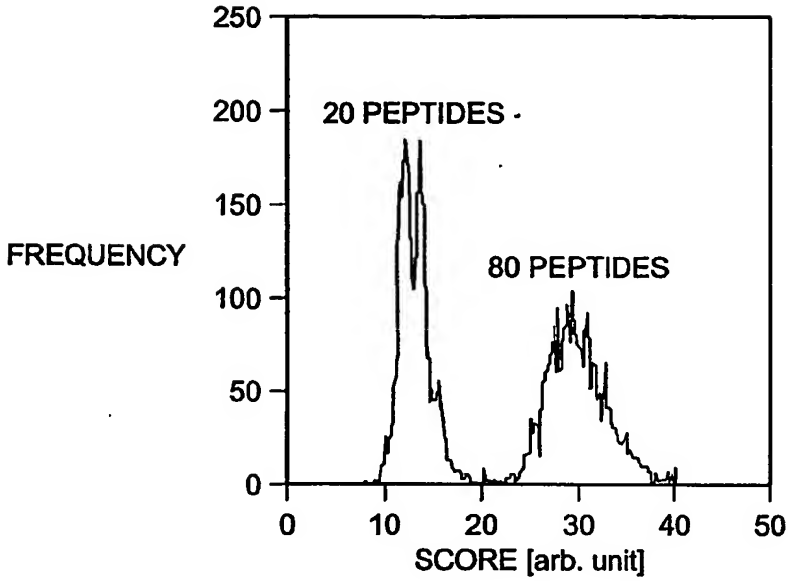

250
200
150
100
50
0
FREQUENCY
20 PEPTIDES
80 PEPTIDES
0
10
20
30
40
50
SCORE [arb. unit]

FIG. 1b

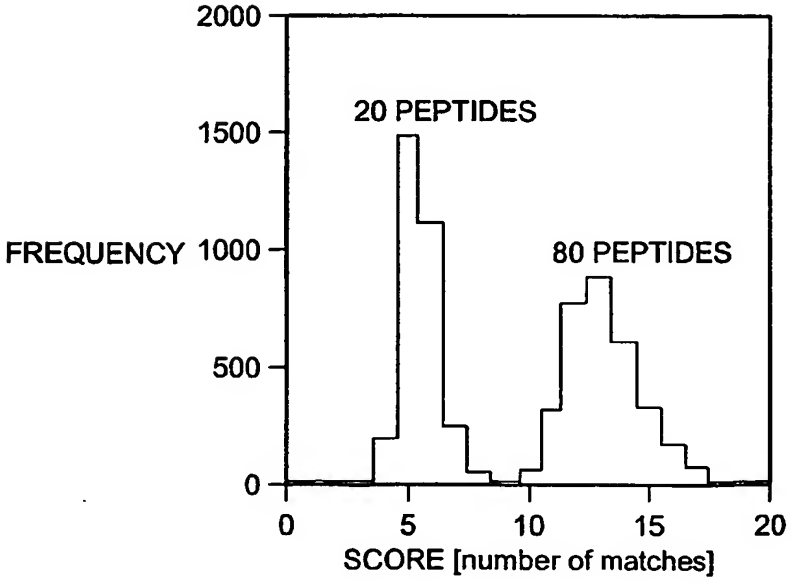

2000
1500
1000
500
0
FREQUENCY
20 PEPTIDES
80 PEPTIDES
0
5
10
15
20
SCORE [number of matches]

FIG. 2a

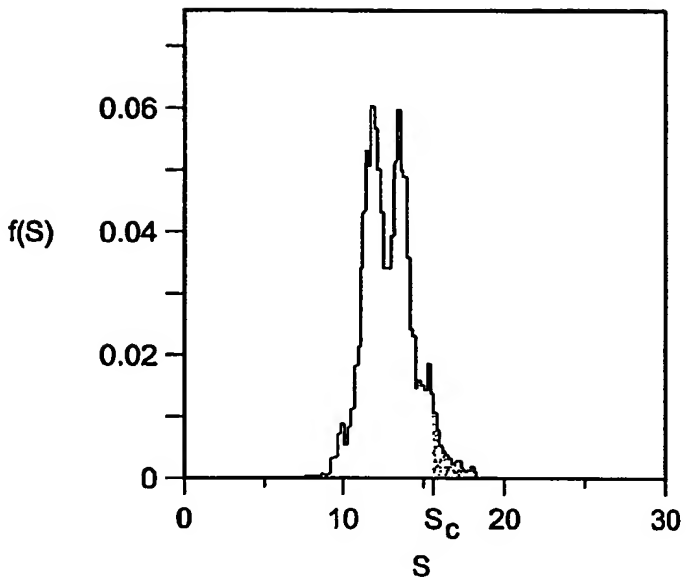

f(S)
0.06
0.04
0.02
0
0
10
S_c
20
30
S

FIG. 2b

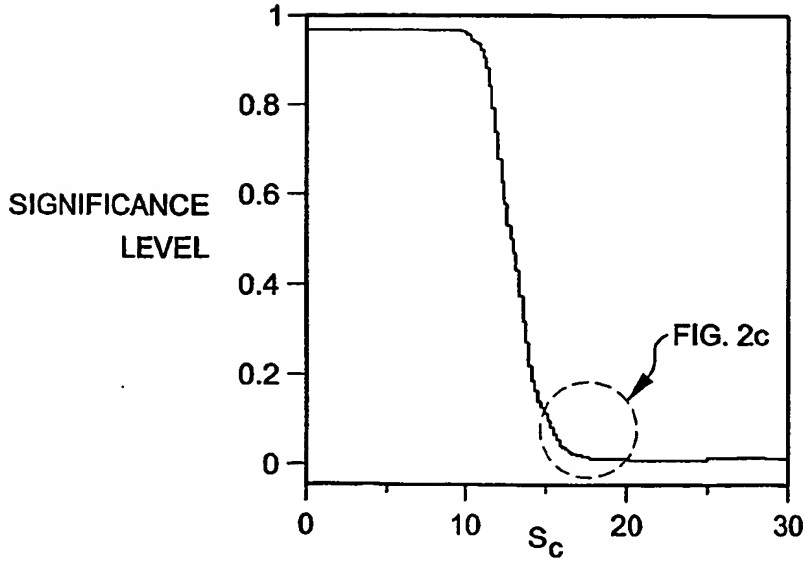

SIGNIFICANCE LEVEL
1
0.8
0.6
0.4
0.2
0
FIG. 2c
0
10
S_c
20
30

FIG. 2c

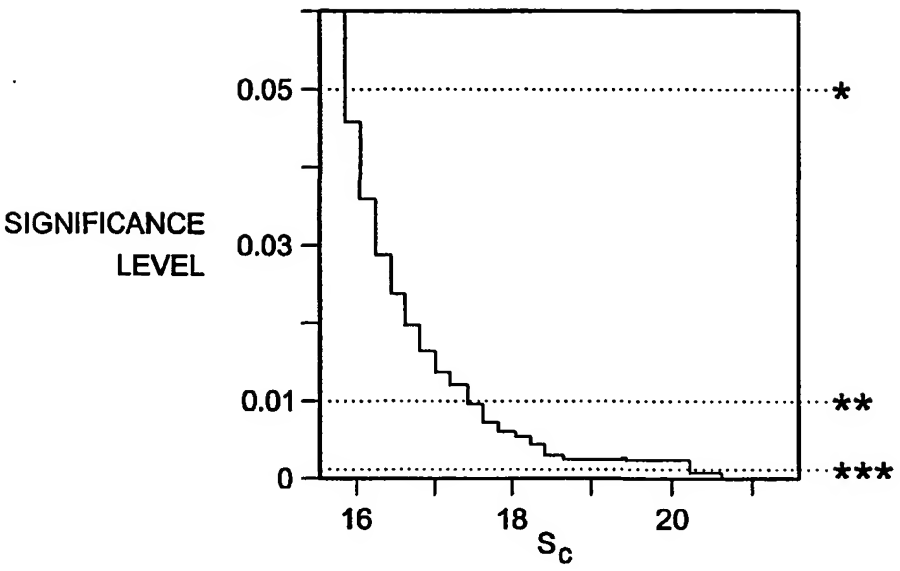

SIGNIFICANCE LEVEL
0.05
0.03
0.01
0
*
**
***
16
18
20
Sc

FIG. 2d

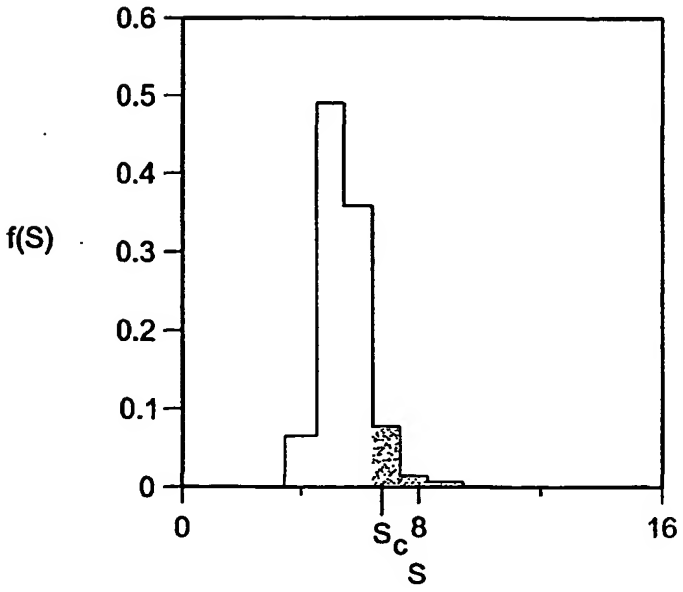

f(S)
0.6
0.5
0.4
0.3
0.2
0.1
0
0
Sc
8
16
S

FIG. 2e

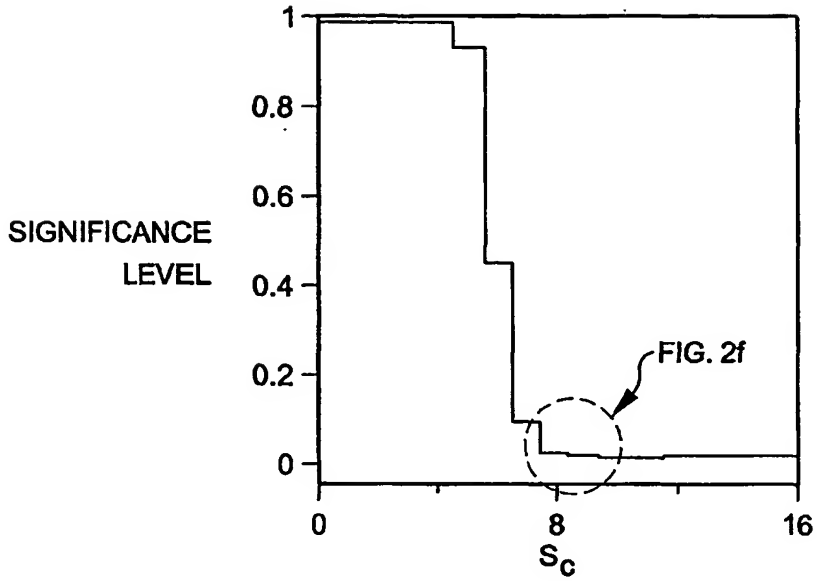

SIGNIFICANCE LEVEL
1
0.8
0.6
0.4
0.2
0
FIG. 2f
0
8
16
S_c

FIG. 2f

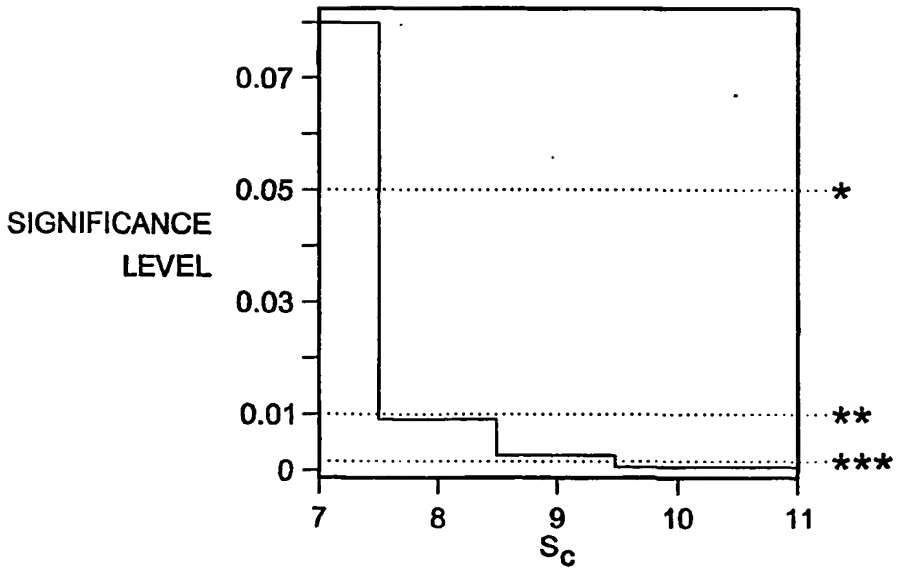

SIGNIFICANCE LEVEL
0.07
0.05
0.03
0.01
0
*
**
***
7
8
9
10
11
S_c

FIG. 3a

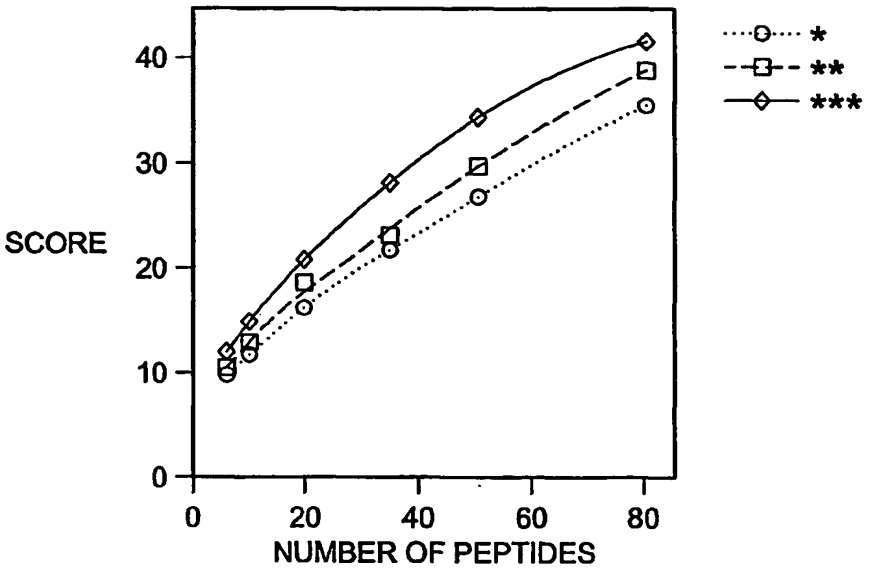

SCORE
40
30
20
10
0
0
20
40
60
80
NUMBER OF PEPTIDES
*
**
***

FIG. 3b

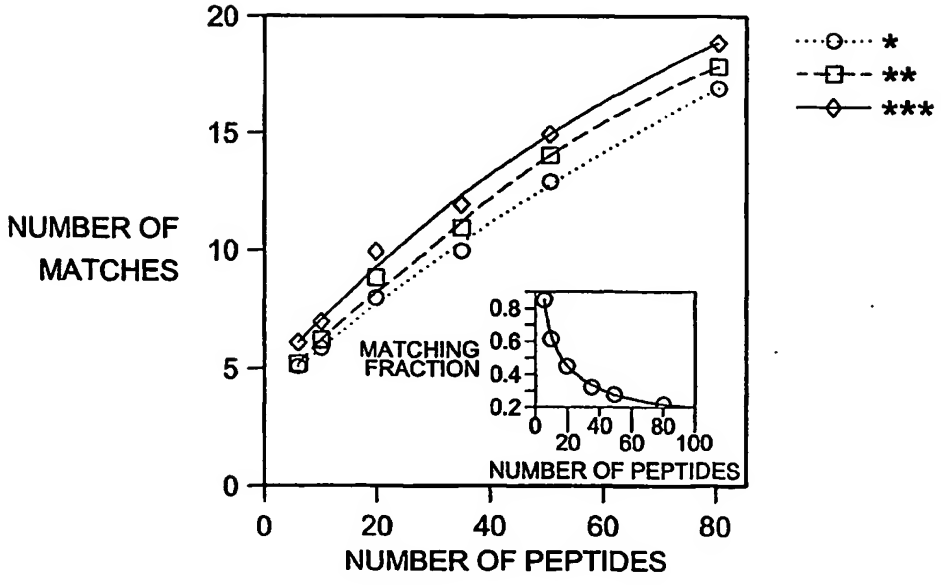

NUMBER OF MATCHES
20
15
10
5
0
0
20
40
60
80
NUMBER OF PEPTIDES
MATCHING FRACTION
0.8
0.6
0.4
0.2
0
20
40
60
80
100
NUMBER OF PEPTIDES
*
**
***

FIG. 4a

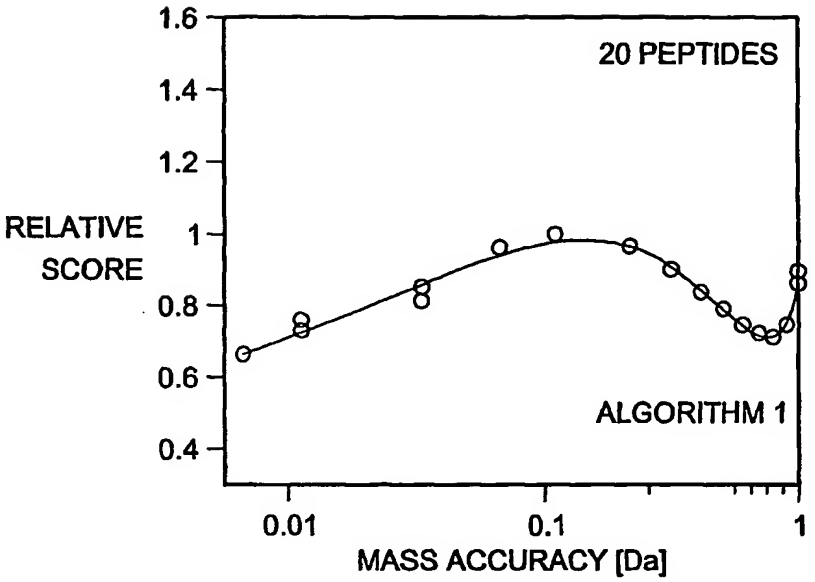

1.6
1.4
1.2
1
0.8
0.6
0.4
RELATIVE SCORE
20 PEPTIDES
ALGORITHM 1
0.01
0.1
1
MASS ACCURACY [Da]

FIG. 4b

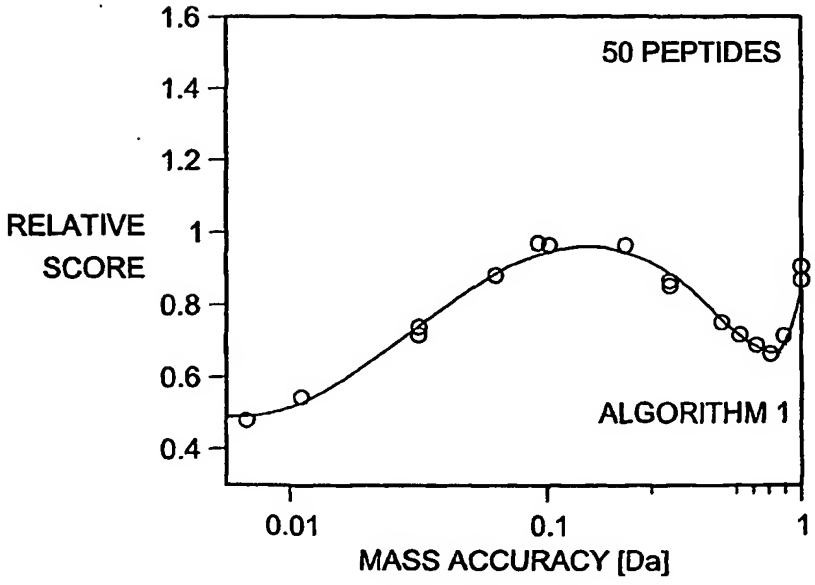

1.6
1.4
1.2
1
0.8
0.6
0.4
RELATIVE SCORE
50 PEPTIDES
ALGORITHM 1
0.01
0.1
1
MASS ACCURACY [Da]

FIG. 4c

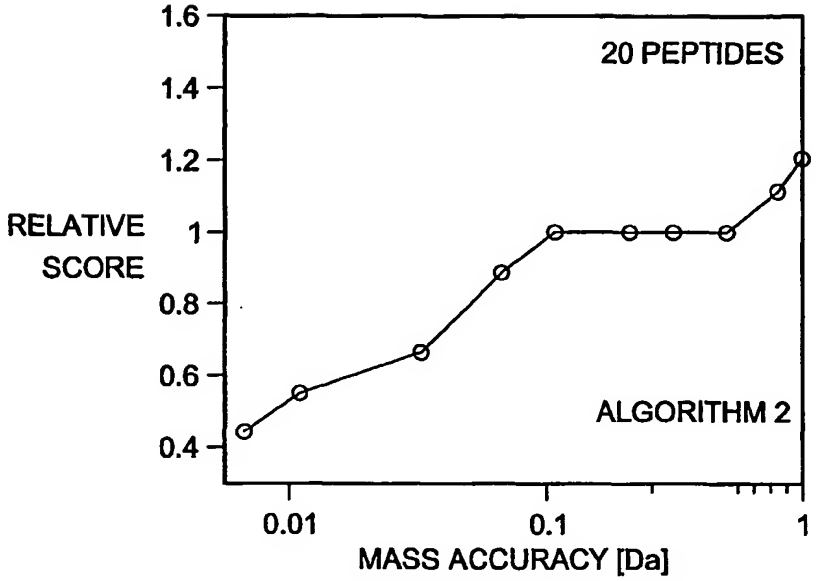

1.6
1.4
1.2
1
0.8
0.6
0.4
RELATIVE SCORE
20 PEPTIDES
ALGORITHM 2
0.01
0.1
1
MASS ACCURACY [Da]

FIG. 4d

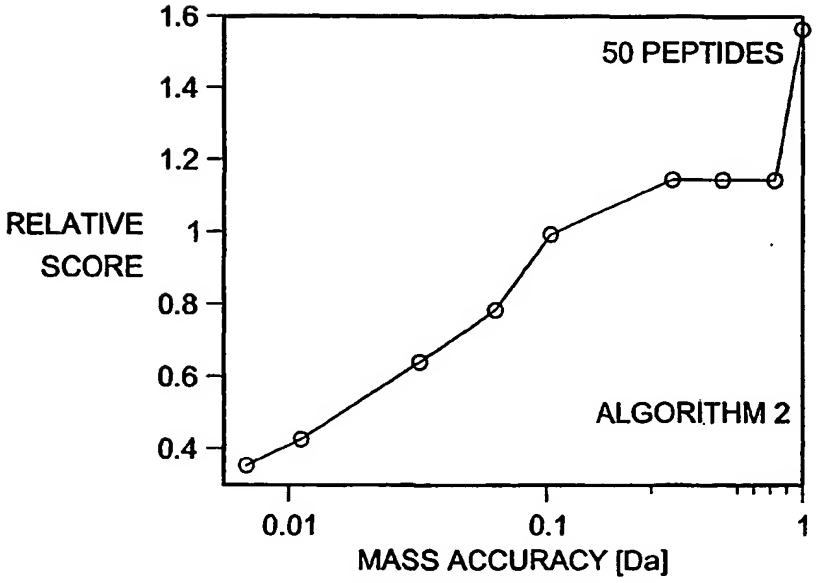

1.6
1.4
1.2
1
0.8
0.6
0.4
RELATIVE SCORE
50 PEPTIDES
ALGORITHM 2
0.01
0.1
1
MASS ACCURACY [Da]

FIG. 5a

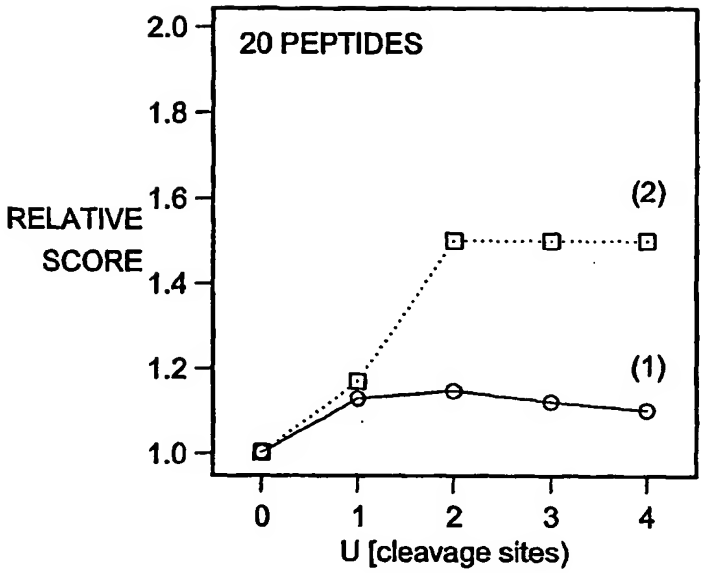

20 PEPTIDES
RELATIVE SCORE
2.0
1.8
1.6
1.4
1.2
1.0
(2)
(1)
0
1
2
3
4
U [cleavage sites)

FIG. 5b

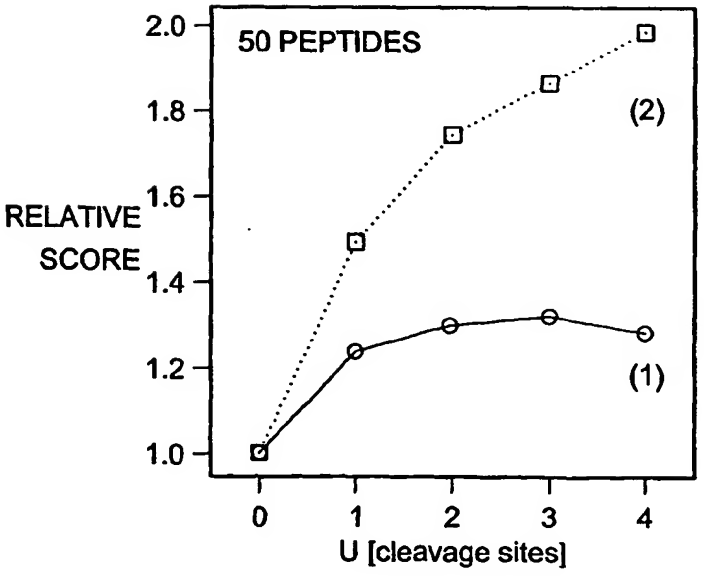

50 PEPTIDES
RELATIVE SCORE
2.0
1.8
1.6
1.4
1.2
1.0
(2)
(1)
0
1
2
3
4
U [cleavage sites]

FIG. 6a

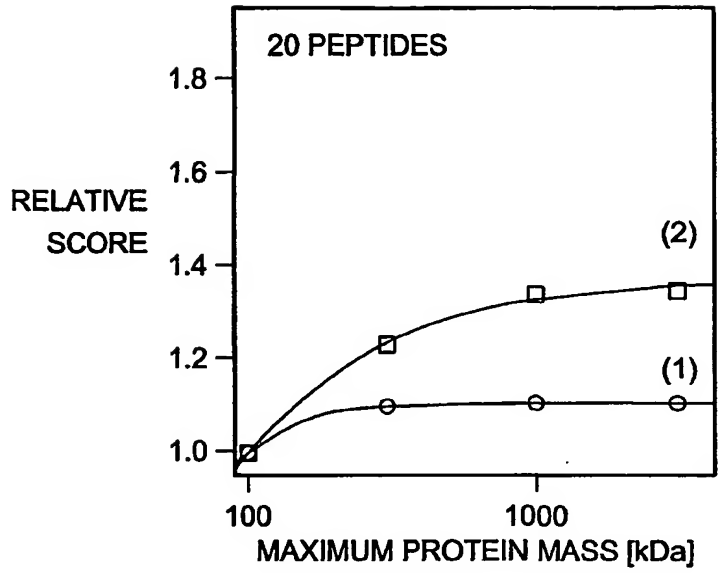

20 PEPTIDES
RELATIVE SCORE
1.8
1.6
1.4
1.2
1.0
(2)
(1)
100
1000
MAXIMUM PROTEIN MASS [kDa]

FIG. 6b

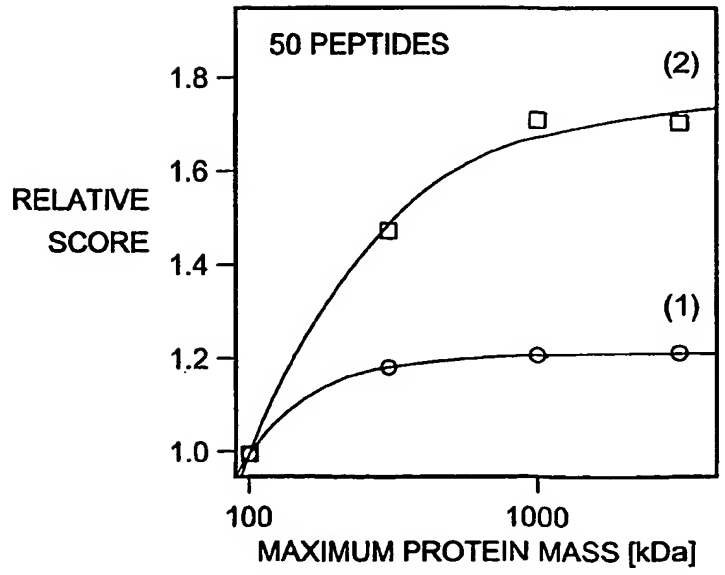

50 PEPTIDES
RELATIVE SCORE
1.8
1.6
1.4
1.2
1.0
(2)
(1)
100
1000
MAXIMUM PROTEIN MASS [kDa]

FIG. 7a

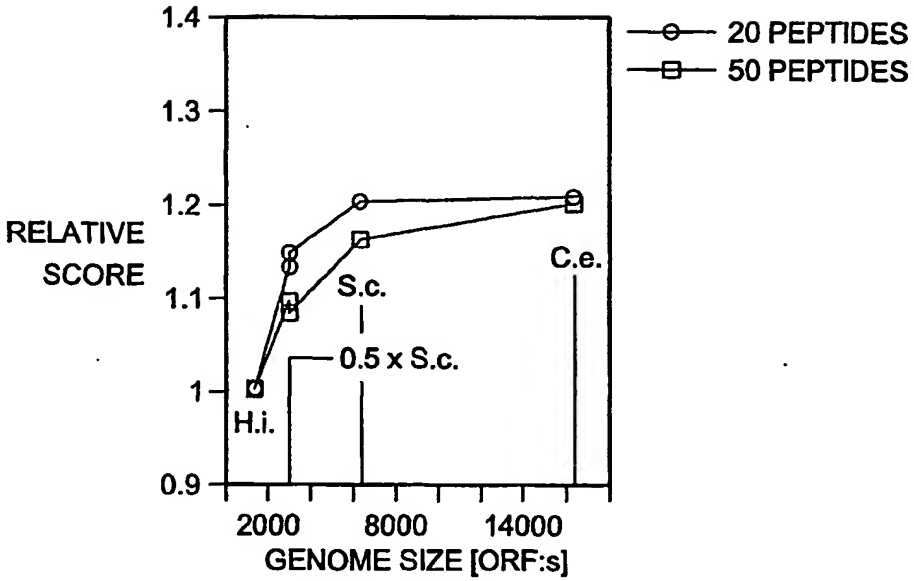

1.4
1.3
1.2
1.1
1
0.9
RELATIVE SCORE
20 PEPTIDES
50 PEPTIDES
C.e.
S.c.
0.5 x S.c.
H.i.
2000
8000
14000
GENOME SIZE [ORF:s]

FIG. 7b

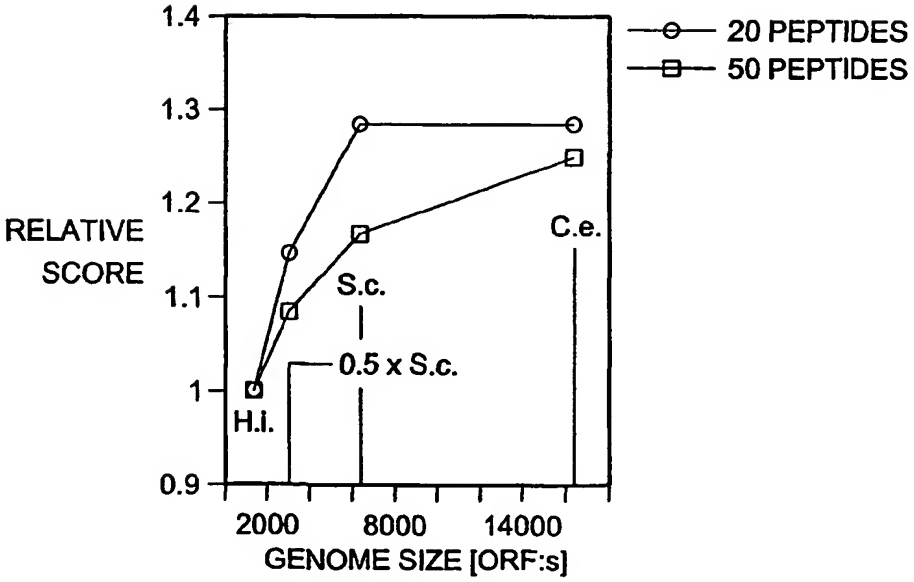

1.4
1.3
1.2
1.1
1
0.9
RELATIVE SCORE
20 PEPTIDES
50 PEPTIDES
C.e.
S.c.
0.5 x S.c.
H.i.
2000
8000
14000
GENOME SIZE [ORF:s]

FIG. 8 PROTEIN IDENTIFICATION DATABASE SEARCH WITH INFORMATION FROM A PEPTIDE MAP

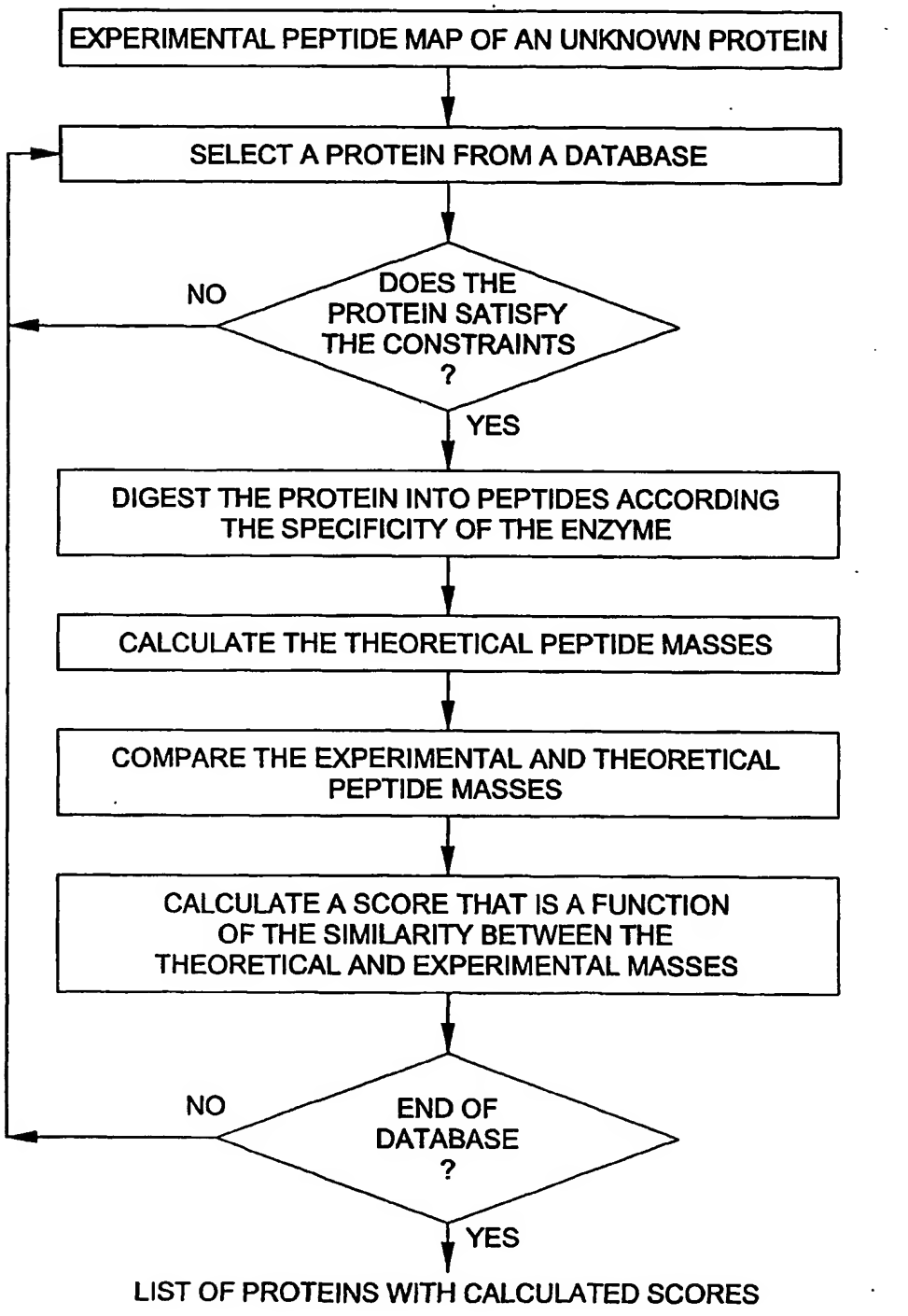


FIG. 9 OBTAINING A SCORE FREQUENCY DISTRIBUTION

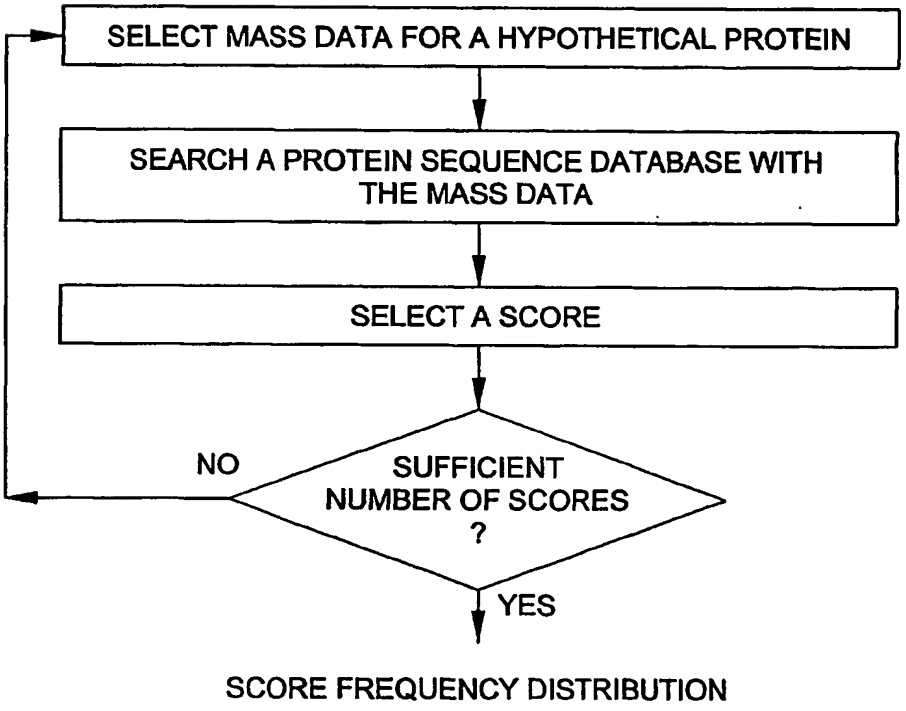


FIG. 10 GENERATION OF A HYPOTHETICAL PEPTIDE MAP

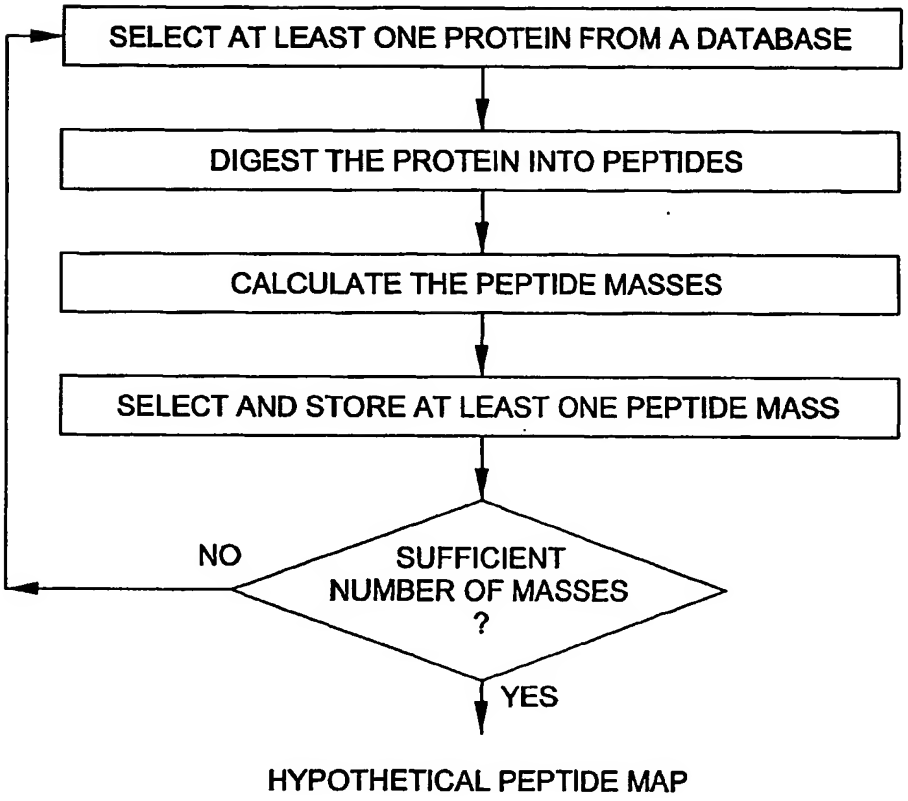
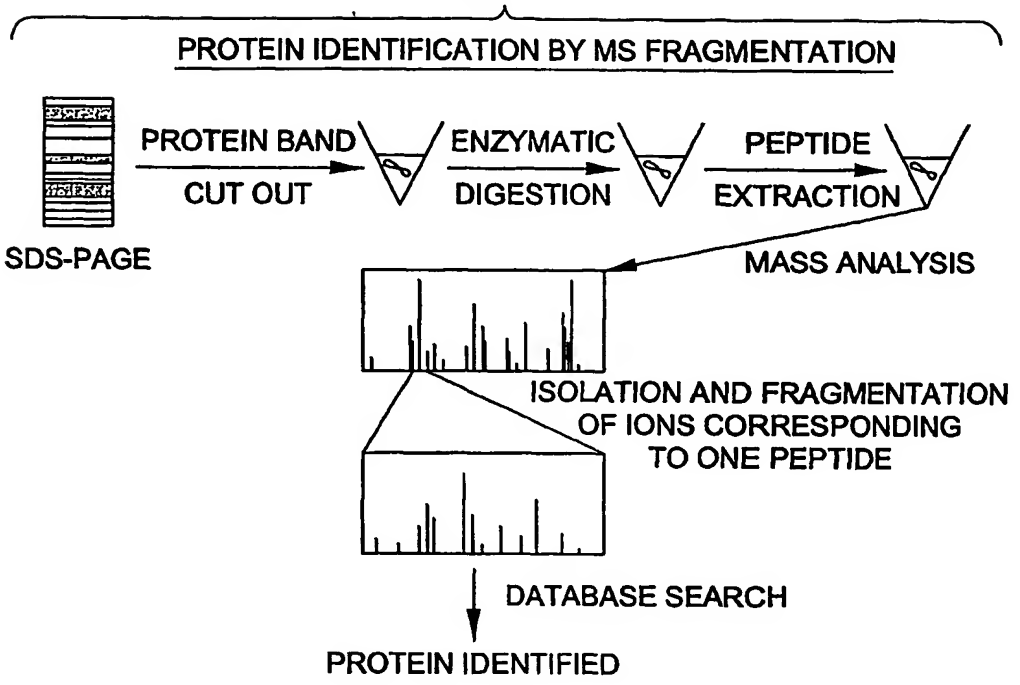


FIG. 11


PROTEIN IDENTIFICATION BY MS FRAGMENTATION
PROTEIN BAND CUT OUT
ENZYMATIC DIGESTION
PEPTIDE EXTRACTION
SDS-PAGE
MASS ANALYSIS
ISOLATION AND FRAGMENTATION OF IONS CORRESPONDING TO ONE PEPTIDE
DATABASE SEARCH
PROTEIN IDENTIFIED

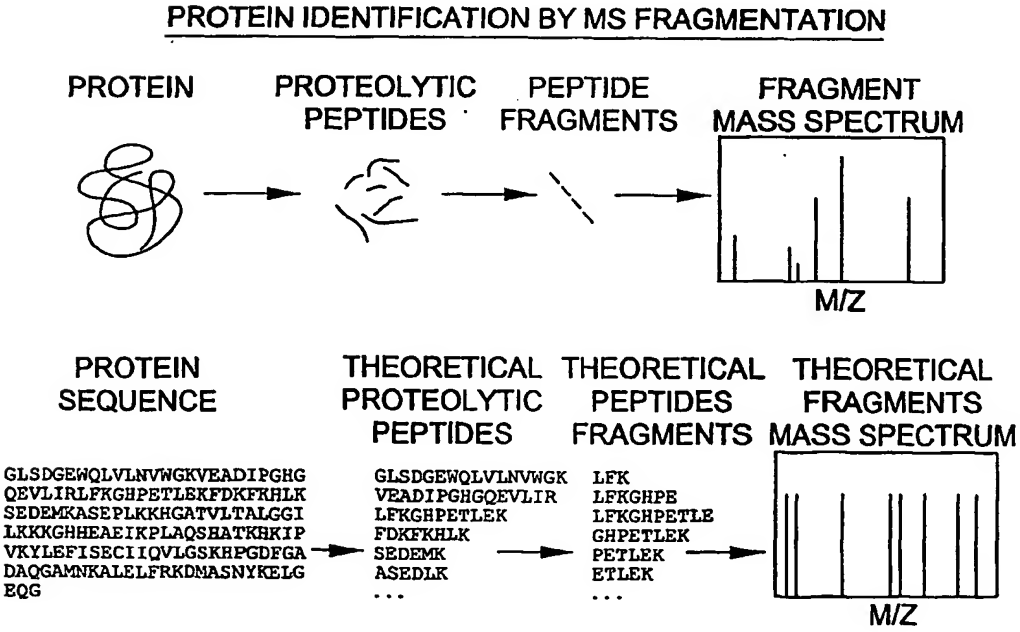
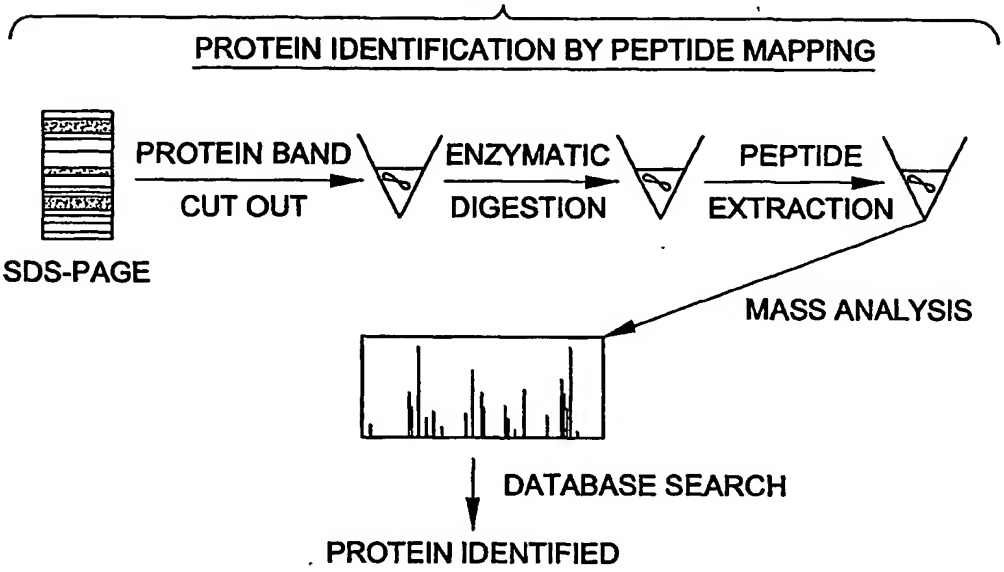

PROTEIN IDENTIFICATION BY MS FRAGMENTATION
PROTEIN
PROTEOLYTIC PEPTIDES
PEPTIDE FRAGMENTS
FRAGMENT MASS SPECTRUM
M/Z
PROTEIN SEQUENCE
GLSDGEWQLVLNVWGKVEADIPGHG
QEVLIRLFKGHPETLEKFDKFKHLK
SEDEMKASEPLKKHGATVLTALGGI
LKKKGHHEAEIKPLAQSHATKHKIP
VKYLEFISECIIQVLGSKHPGDFGA
DAQGAMNKALELFRKDMASNYKELG
EQG
THEORETICAL PROTEOLYTIC PEPTIDES
GLSDGEWQLVLNVWGK
VEADIPGHGQEVLIR
LFKGHPETLEK
FDKFKHLK
SEDEMK
ASEDLK
...
THEORETICAL PEPTIDES FRAGMENTS
LFK
LFKGHPE
LFKGHPETLE
GHPETLEK
PETLEK
ETLEK
...
THEORETICAL FRAGMENTS MASS SPECTRUM
M/Z

FIG. 12


PROTEIN IDENTIFICATION BY PEPTIDE MAPPING
PROTEIN BAND
CUT OUT
ENZYMATIC
DIGESTION
PEPTIDE
EXTRACTION
SDS-PAGE
MASS ANALYSIS
DATABASE SEARCH
PROTEIN IDENTIFIED

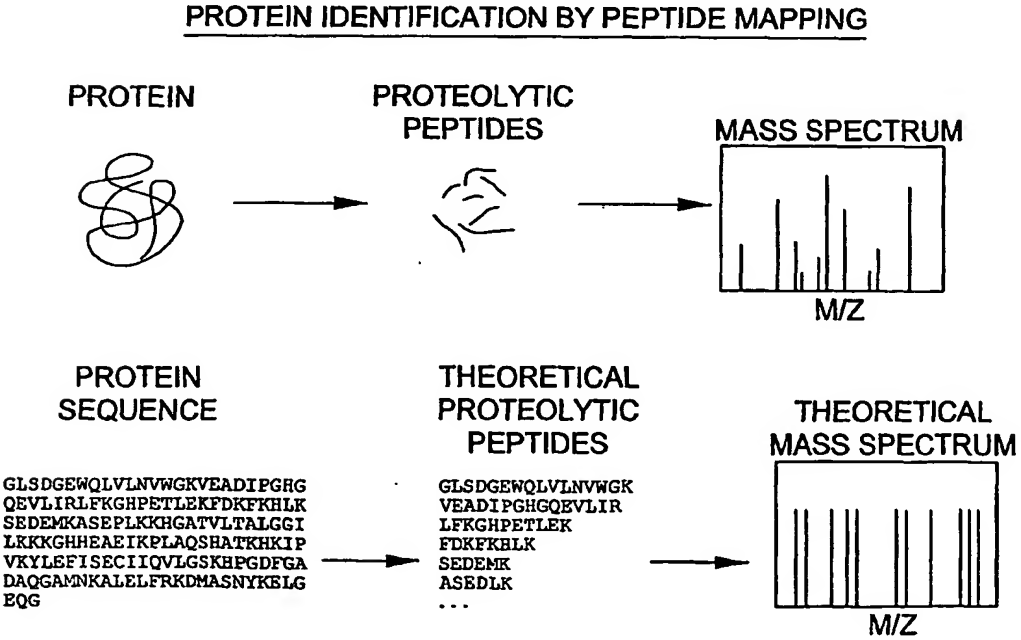

PROTEIN IDENTIFICATION BY PEPTIDE MAPPING
PROTEIN
PROTEOLYTIC PEPTIDES
MASS SPECTRUM
M/Z
PROTEIN SEQUENCE
GLSDGEWQLVLNVWGKVEADIPGHG
QEVLIRLFKGHPETLEKFDKFKHLK
SEDEMKASEPLKKHGATVLTALGGI
LKKKGHHEAEIKPLAQSHATKHKIP
VKYLEFISECIIQVLGSKHPGDFGA
DAQGAMNKALELFRKDMASNYKELG
EQG
THEORETICAL PROTEOLYTIC PEPTIDES
GLSDGEWQLVLNVWGK
VEADIPGHGQEVLIR
LFKGHPETLEK
FDKFKHLK
SEDEMK
ASEDLK
...
THEORETICAL MASS SPECTRUM
M/Z

FIG. 13 FLOW CHART SHOWING THE STEPS IN A COMPLETE PROTEIN IDENTIFICATION MS/MS EXPERIMENT

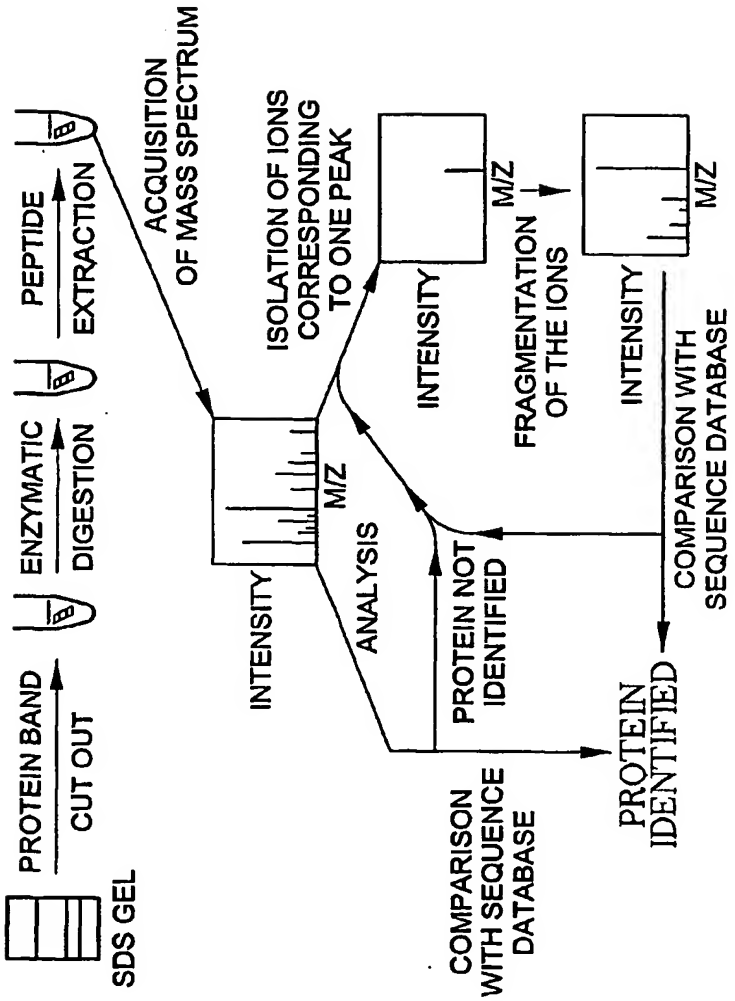
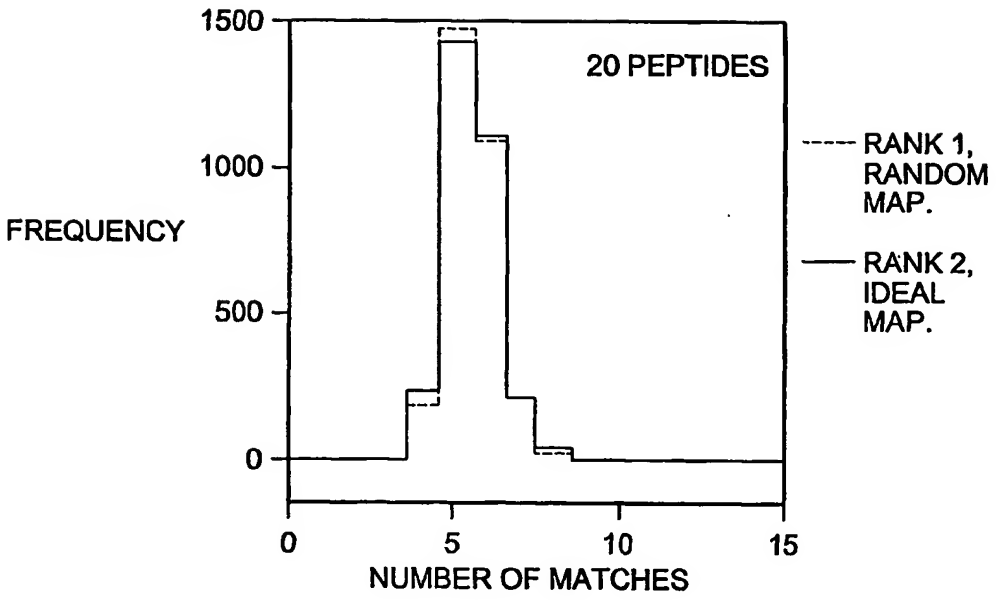


FIG. 14


1500
1000
500
0
FREQUENCY
20 PEPTIDES
RANK 1, RANDOM MAP.
RANK 2, IDEAL MAP.
0
5
10
15
NUMBER OF MATCHES

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/16638

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G06F 19/00
US CL :702/19

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 702/19

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

USPAT, Medline, Biosis, Derwent WPI, CAPlus
search terms: peptide, mass spectrometry, frequency distribution

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CORDWELL et al. Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. Electrophoresis. 1995, Vol. 16, pages 438-443, especially page 442.	18, 47, 48
X	COTTRELL et al. Protein Identification by Peptide Mass Fingerprinting. 1994, Vol. 7, pages 115-124, especially page 117.	1, 8, 13, 14, 16, 20-22, 26, 36
Y		18, 47, 48

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"B" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 AUGUST 2000	04 OCT 2000

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	JOHN S. BRUSCA
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/16638

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.
1, 8, 13, 14, 16, 18, 20, 21, 22, 26, 36, 46, 47

Remark on Protest
☐ The additional search fees were accompanied by the applicant's protest
☐ No protest accompanied the payment of additional search fees

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/16638

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

This application contains claims directed to more than one species of the generic invention. These species are deemed to lack Unity of Invention because they are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for more than one species to be searched, the appropriate additional search fees must be paid. The species are as follows:

Methods of generating frequency distribution scores of mass data of:
1) proteins
2) nucleic acids
3) polysaccharides

The claims are deemed to correspond to the species listed above in the following manner:

1) protein species: 8, 16, 18, 21, 26, 46, 47
2) nucleic acid species: 9, 27
3) polysaccharide species: 10, 28

The following claims are generic: 1, 14, 13, 20, 22, 36

The species listed above do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, the species lack the same or corresponding special technical features for the following reasons: The claims are drawn to methods of generating data of mutually exclusive types of molecules and therefore do not share a special technical feature.